

**Alternative Assessment:
Research, Resources, and Use Guidelines**

**Prepared for:
Arizona Department of Education,
Career and Technical Education Division
Project # ED02-0055**

Carol A. Norris, Ph. D.

January 30, 2003

Executive Summary

This research project was undertaken on behalf of the Arizona Department of Education, Career and Technical Education Division, to examine materials and to develop:

- 1) criteria and guidelines for alternative assessment materials/tools;
- 2) assessment recommendations for the CTE Cross-Program Competencies;
- 2) rubrics for judging and identifying appropriate alternative assessment tools.

The focus of the research is on standards/criteria for educational accountability with particular emphasis on alternative assessment methodologies to ascertain Level III Career and Technical Education program competency attainment, including nine Cross-Program Competencies recently adopted by the CTE Division. Other purposes include to examine and recommend possible locally-developed materials which could be adopted/adapted for statewide use, to identify resources, and to draft administrative guidelines to implement alternative assessment programs.

Assessment strategies and test types were reviewed to use as benchmarks for student performance assessment and accountability reporting. Throughout this report a differentiation is made between accountability and assessment models. *Accountability* is used broadly to mean any test/measurement activity (usually a paper/pencil test) in which student scores are grouped and used as the basis for reporting attainment for federal, state, or local reports to prove educational impact.

Assessment is used broadly to mean instructional testing in which assessments are designed to help teachers improve student learning. The term incorporates three commonly used terms including alternative, authentic, and performance assessment. The report concentrates on effective assessment strategies to utilize within a continuous assessment program, perhaps culminating with an end-of-term/exit exam of some type. Such exams are typified in industry and vendor-specific certifications or endorsements applicable to career/technical education programs.

Vocational/technical education goals have not changed significantly in the past 20 years. What has changed is available technology and its effect on the nature of work, the organizational structure of school systems, and the development of instructional standards and curriculum modifications to reflect those standards. In addition, student performance expectations/criteria and the use of formal accountability systems documenting performance has become commonplace. Accountability emphasizes performance standards (how good is good enough) and student performance (how close are students to meeting the standard).

Throughout the report a pro/con writing approach has been used to describe issues related to various assessment practices, standards, and appropriate uses, pitfalls and criteria for selection of respective test types. Comprehensive resource and reference materials and reprints of professional/technical articles are included in the Appendix. Finally, a discussion of professional development considerations and suggested local district/state education agency implementation strategies for alternative assessment are detailed.

Special needs population assessment issues and issues related to competency tracking and database reporting systems are reviewed in a limited manner. No attempt has been made to

delve into the technical requirements for effective, reliable, valid, and appropriate test item construction because it is a highly technical area.

Several recommendations address professional development activities in this area.

Arizona Academic Standards include eight (8) Workplace Skills. In this report, Workplace Skills are separated when referencing Academic Standards. This is only for ease of comparison with other states/entities and is not to suggest that they are not appropriately part of the Academic Standards.

With reauthorization of the Elementary and Secondary Education Act (No Child Left Behind Act), the federal government has required all states to establish accountability systems that include high-stakes testing. This has led states, local districts, textbook publishers, and other providers of content for education to begin work aligning standards, curriculum, and assessments. Arizona has done so by aggressively updating its' Curriculum Frameworks, including identifying appropriate possible business/industry/vendor certifications to validate student attainment.

Another consideration has been to address mandatory performance indicator information for state accountability reporting purposes. Reporting requirements have focused attention on the problems associated with documenting and tracking student performance and the use of powerful electronic systems to provide administrative support and performance analysis information.

The issue of *how to measure* and report student performance centers on the availability and acceptance of a variety of performance measures, including state-mandated performance assessment(s) in either/both academic, workplace, and vocational competencies, including core skills. Although there is evidence that generic skills are being taught in schools, there is great ambiguity about what they are and about how they are integrated in the CTE curriculum.

Summary/Conclusions

This report addresses a variety of public/professional testing issues, test types and uses, performance measures and accountability reporting, district assessment and data management capabilities, recently adopted CTE Cross-Program Competencies, potential assessment approaches, and student assessment practices.

Essentially, there are six questions to resolve as the State moves toward improving its accountability system, districts move closer to meeting state performance measures, and student assessment practices are revised/improved. These questions include:

- 1) How do student performance assessment practices continue to improve the learning environment and simultaneously accomplish accountability and state performance measures reporting?
- 2) What steps can be accomplished to resolve local district/state issues that surround accountability versus assessment practices?

- 3) How do districts implement and absorb costs associated with using mandated industry/vendor standards and certifications for CTE student assessment and State reporting purposes?
- 4) What are the appropriate uses of various test types and are they acceptable as “in lieu of” tests for state accountability and performance measures reporting?
- 5) How can Arizona implement an assessment and performance reporting system that is acceptable to both local districts and the State Department of Education, Career Technical Education Division?
- 6) Should a common assessment for Work Place Skills and Cross-Program Competencies be developed or should the State develop test items to include these skills and competencies in the AIMS test?

Conclusions

- Organizations can be divided and grouped into similar categories: Assessment/Test providers, Curriculum developers (with some assessment capabilities potentially), and Standards developers.
- There is limited commonality in materials prepared/available from consortiums, states, organizations, and the U. S. Department of Education with regard to performance standards, academic/vocational curriculum integration models, and assessment strategies.
- The SCANS competency areas are still widely regarded as a standard for assessing workplace readiness skills, although not all states and consortiums have adopted the SCANS skill areas.
- Some States (Oregon, New York, Arkansas, Ohio, Oklahoma, Wisconsin) have made significant strides to integrate academic and technical skill standards into common curriculum goals with career-related learning standards. Use of alternative assessments are a component of these initiatives.
- Common (core/cross-program) competencies are beginning to be articulated by varying States and consortiums, although most can't/won't articulate them specifically, unless one's state is a member of the pilot group or consortium.
- Career Clusters are also regarded as another set of standards around which curriculum and assessments are organized. Significant data is available from <http://www.careerclusters.org>

- According to the Education Commission of the States¹ Maryland is the only state that has established World of Work and Survival Skills for inclusion in the state's core assessment system.
- Two of the "cross-program competencies" adopted by Arizona are not reflected in the SCANS and/or CTE Cross-Program Competencies (namely Evaluate the role of small business in the economy and Business/financial management for entrepreneurs), nor are they typically assessed for in the myriad of assessment options available through testing sources and consortiums.
- Two of the Arizona Cross Program Competencies (Develop an individual career plan and Evaluate the role of small business in the economy) do not cross-reference to either SCANS or the Arizona Work Place Skills.
- "Develop an individual career plan" is part of the process to "Prepare for employment" and could be combined with it, rather than continued as a separate Cross-Program Competency for Arizona programs.
- Based on limited information, it appears that "Evaluate the role of small business in the economy" as a separate Cross-Program Competency should be reconsidered.
- Review of the crosswalk clearly demonstrates that the Academic Standards essentially already include the CTE Cross-Program Competencies. Assuming the crosswalk satisfies criteria of "adequacy and accuracy," there is little, if any, need to develop additional competencies/indicators.
- There is no standard practice regarding the use of skill certificates and industry credentials for secondary and community college vocational/technical education students.
- Districts are generally unwilling to replace their existing assessments with a State model and are more willing to select an alternative assessment model acceptable to both them and the State in order to maintain "approved program" status for accountability purposes in Arizona.
- There are wide differences among the states in the degree to which Cross-Program Competencies, Workplace Skills, and Technical Standards are included/excluded from a particular state's core assessment and reporting system(s).
- There is little commonality among Arizona school districts with respect to performance assessment systems, test type utilization, administrative guidelines, and testing practices.

¹ ECS Clearinghouse Notes, Advanced Placement Courses and Examinations. (January 2000). Education Commission of the States, Denver, CO.

- Database capabilities to track and report student performance and to respond to ADE performance indicators differs significantly among Arizona districts ranging from hand posting to sophisticated electronic data management.
- Some districts purchase tests/test services, but per student costs have prohibited many districts from purchasing such industry/vendor-prepared assessments.
- The options of developing an Arizona test, adopting another state's test(s) or industry-endorsed tests, and/or purchasing vendor testing services has not been resolved for CTE technical skills assessment.
- The Arizona AIMS test does not include Workplace Skills, although they are included in the Arizona Standards.
- The Arizona Workplace Skills, SCANS Skills, and all but two CTE Cross-Program Competencies are closely aligned and they, in turn, are adequately included in the Arizona Academic Standards
- Integrated academic and vocational/technical curriculum practices in Arizona have the potential to improve academic and workplace skills development.
- No single test type emerges as "most preferred" for vocational/technical assessments, although performance-based assessment is widely used in Arizona and other states.
- District-level professional resources to assist teachers in developing test items and test administration skills are available in limited manner.
- Test development and test item writing skills require higher levels of technical competency than may exist in some school districts.
- Many districts retain on-site staff and expend large amounts of personnel, time, and money to keep reporting requirements and instructional content and assessment practices current through professional development activities.
- Because of group rather than individual scoring practices, team and chapter CTSO events should not be considered for individual student vocational competency assessment and accountability reporting.
- It is questionable that individual CTSO event content, judges' selection practices, event administration guidelines, and scoring rubrics are cross-referenced to the respective instructional program(s), cross-program competencies, and/or work place skills to determine vocational competency assessment.
- Recently implemented Design Team makeup, procedures, and content requirements in the ADE Curriculum Frameworks provide considerably more industry/vendor

assessment and certification resource information for Arizona CTE teachers than in past curriculum guides.

- New Curriculum Frameworks follow, for the most part, common formats; however, uncommon formats have been utilized when referencing available assessment and certification sources.

Recommendations

Thirty recommendations addressing assessment practices, accountability tracking and reporting system capability, state accountability issues, test design types and technical skill competencies, professional development needs, and alternative assessment implementation strategies are included in the report for consideration.

Additional information or full copies of the report may be obtained from Ms. Marilee Johnson or Ms. Paulette Regan at the Arizona Department of Education, Career and Technical Education Division.

Acknowledgments

There were many contributors who provided resource materials, time, energy, and guidance for this report. Their professional expertise was invaluable. My thanks to the following resource persons for their generous sharing of materials and input for the project:

Resource Team members: Gay Evangelista (Peoria Unified District), Georgia Merrick (Tempe Union District), Lois Lamer (Mingus Union High School District), Dennis Fiscus (Arizona Department of Education), Harold Kirchner (Yuma Union High School District), Jon Linberg (Kingman Unified District), and Tony Maldonado (Mesa Unified School District)

Review Team members: Delores Watkins (Apache Junction Unified District), Marilyn Ruggles (Tempe Union District), Mark Hamilton (Gilbert Unified District), James Brown (Peoria Unified District) and Barry Williams (Round Valley Unified District)

Rubrics: “Rubrics Madness” Judy Balough (Arizona State University), Michelle Crary (Desert Vista High School), Chris Libette Garcia (Metro Tech High School), Sue Crumrine (Winslow High School) and Nanette Gillispie (Peoria School District)

Jon Lindberg (Tech Prep Coordinator and Vocational Director, Kingman Unified District) and Kathy Prather (Tucson Unified School District)

Licenses/certifications: Research Report 10/15/01; Dr. Jack Elliot and Ms. Denise Davies (University of Arizona, Agricultural Education)

Academic Standards, Cross-Program Competencies, and Work Place Skills Crosswalk: Susan Cooper (Northern Arizona University, Institute for Future Workforce Development)

Research assistance: Jamie Rondeau (industry consultant) and Vaughn Croft (education consultant)

On-line vocational competency reporting: Georgia Merrick and Marilyn Ruggles (Tempe Union High School District), Dean Peterson (Glendale Union High School District) and Delores Watkins (Apache Junction Unified District)

Course competency matrices and competency assessment information: Mountain Pointe and McClintock High Schools, Georgia Merrick, and Marilyn Ruggles (Tempe Union High School District)

Performance-based assessment booklets: Dean Peterson, Vocational Director and Darlene Benford, Administrative Assistant (Glendale Union High School District)

Videos: “Career & Technical Education Works” and “Academics In Action” Gay Evangelista (Peoria Unified School District)

Arizona Department of Education Project Director: Mrs. Marilee Johnson

Arizona Department of Education Project Liaison: Ms. Paulette Regen

Table of Contents

TOPIC	PAGE
Introduction.....	1
Methodology.....	2
Limitations.....	2
Overview of CTE Programs.....	3
Accountability and Program Performance Systems.....	4
Rationale to Support or Reject High-Stakes Testing.....	6
State Accountability Reporting.....	9
Performance Indicators for CTE Programs.....	12
General Purposes of Student Assessment.....	14
Assessment Principles.....	16
Standardized Tests.....	19
Criterion-Referenced Tests.....	22
Standards and Assessment.....	27
Cross-Program Competencies and Workplace Skills Assessments.....	40
Crosswalk: Arizona Academic Standards and CTE Cross-Program Competencies.....	43
National Standards and Assessments for CTE Programs.....	44
Assessment Types and Instructional Uses.....	61
Rubrics and Checklists.....	78
CTSO Events: A Possible Assessment Vehicle.....	95
Professional Development.....	96
Alternative Assessment Implementation Strategies.....	98
Draft ADE Approval Process for Local District Prepared Assessments.....	100
Conclusions.....	100
Recommendations.....	103
References.....	106
Web and Rubrics Resources.....	112
 Glossary	
General Educational Terms.....	121
Harcourt Brace Tests and Measurement Terms.....	125

List of Tables

Table I: Core Indicators and 2003 Performance Measures	13
Table II: Comparison of SCANS, Arizona Workplace Skills and CTE Cross-Program Competencies	37
Table III: Arizona, Virginia, and New York Cross-Program Competencies	39
Table IV: Cross-Program and Workplace Skills Assessment Sources	40
Table V: Arizona Licensed Occupations.....	51
Table VI: Arizona Apprenticeship Programs.....	54
Table VII: Academic and Technical Skills Assessment and Certification Sources.....	57
Table VIII: Authentic Assessment Tools/Performance Activities.....	76

LIST OF APPENDICES

Appendix F: Articles: “Alignment of Standards and Assessments as an Accountability Criterion” and “Fundamental Assessment Principles For Teachers and School Administrators”	134
Appendix G: Articles: “Authentic Assessment-Basic Definitions and Perspectives,” “Authentic Assessment Tools,” “Academic Testing Test Design and Construction,” “Improving Your Test Questions,” “Item Writing Guidelines,” “Writing Multiple-Choice Test Items,” More Multiple-choice Item Writing Do’s and Don’ts”	148
Appendix H: Articles: “Scoring Rubrics: What, When and How,” and “Designing Scoring Rubrics for Your Classroom”	208

Introduction

This research project was undertaken on behalf of the Arizona Department of Education, Career and Technical Education Division, to examine existing materials and develop:

1. criteria and guidelines for alternative assessment materials/tools
2. alternative assessment materials/tools for the CTE cross-program competencies
3. rubrics for judging and identifying appropriate alternative assessment tools

The focus of the research and materials reviewed is on developing standards/criteria for educational accountability with particular emphasis on alternative assessment methodologies to ascertain Level III Career and Technical Education program competency attainment, including nine Cross-Program Competencies recently adopted by the CTE Division. Other purposes include to examine current assessments, recommend possible locally-developed materials which could be adopted/adapted for statewide use, identify assessment resources, and draft administrative guidelines to implement alternative assessment programs.

Five or more assessment strategies were reviewed to use as benchmarks for student performance assessment and accountability reporting and include:

- industry-validated/commercially prepared tests (such as NSSB and NOCTI)
- true/false and multiple-choice test types
- performance-based assessments, scenarios, and observation checklists
- portfolios and exhibit projects
- test-item banks

Throughout this report a differentiation is made between accountability and assessment models. Accountability is used broadly to mean any test/measurement activity (usually a paper/pencil test) in which student scores are grouped and used as the basis for reporting attainment for federal, state, or local reports to prove educational impact.

Assessment is used broadly to mean instructional testing in which assessments are designed to help teachers improve student learning. The term incorporates three commonly used terms including alternative, authentic, and performance assessment. Any assessment practice or tool that is different from traditional practice is termed an alternative assessment. Thus, paper-and-pencil tests are considered traditional and are one component of the assessment process, but are not considered “alternative.” Authentic and performance assessment terms are also sometimes used interchangeably. Any assessment activity that is contextual is considered authentic; thus, performance-type assessment is one form of authentic assessment.

Two glossaries are included in the report. One contains general terms applicable to assessment and accountability. The second glossary contains specific terms applicable to tests and measurement and provides technical information regarding test item construction, validation, test score interpretation, and presentations techniques.

Methodology

Research was conducted through web searches, email and phone interviews, personal contacts, and review of existing Arizona and national resource materials. The following list summarizes primary contacts and materials utilized for the project:

- Project Resource and Review Team members from local districts for input and guidance;
- Industry groups, test developers/testing services, vendors, educational consortiums, state departments of education, universities, and local district personnel;
- Arizona State Supervisors for Career/Technical Education programs having career/technical student organizations (i.e. FCCLA, FFA, FBLA, etc.);
- ADE curriculum design projects and Academic, Workplace Skills and Cross-Program Competencies;
- Local school district assessment materials and practices;
- Tests/measurements principles and test construction guidelines;
- Arizona-based research, particularly the University of Arizona assessment report (2001) and the Arizona State University rubrics workshop (2002) materials.

Limitations

Special needs population assessment issues and strategies are reviewed in a limited manner. Test administration guidelines and one sample rubric to use with this population is presented. However, there are many resource materials and sample tests that special needs instructors and test administrators could review to broaden their understanding of how to develop assessment instruments, structure, and conduct special needs student assessment activities.

Discussion of some of the issues related to competency tracking and database reporting systems is included in a limited manner. A short description of how three Arizona districts administer their assessment program and database is provided; however, technical requirements, staffing, and implementation strategies for database systems development are not included in this report.

Tests and measurements is a highly specialized, technical discipline. For this reason, a series of technical assistance articles are included in Appendix G to compliment narrative portions of this report. No attempt has been made to delve into the technical requirements for effective, reliable, valid, and appropriate test item construction.

Arizona Academic Standards include eight (8) Workplace Skills. In this report, Workplace Skills are separated when referencing Academic Standards. This is only for ease of comparison with other states/entities and is not to suggest that they are not appropriately part of the Academic Standards.

Overview: Career and Technical Education (CTE) Programs

Since 1917 with the Smith-Hughes Act, national legislation has endorsed and supported career and technical education programs. Subsequent legislation has restated the premise that these programs can help the nation achieve economic vitality and prepare/improve the workforce. Current legislative goals in the Carl D. Perkins Vocational Applied Technology Education Act (PL101-392) of 1990 are to support career/technical programs with emphasis on:

- Academic attainment
- Skill attainment
- Program completion
- Employment or continuing education
- Retention in the job
- Nontraditional participation & completion

Dr. Ernest Boyer, US Department of Education, Office of Adult and Vocational Education, commented “the purpose of education is to empower individuals to live with competence in their community ...to ensure that students acquire the skills and knowledge they need.”² His quote typifies the basis of career and technical education programs. The purposes of high school career and technical education programs today are identified as:

1. Providing career exploration and planning
2. Enhancing academic achievement and motivation to learn more
3. Acquiring generic work competencies and skills useful for employment
4. Establishing pathways for continuing education and lifelong learning³

The National Center for Education Statistics report⁴ defined vocational education as a sequence of courses designed to prepare students for an occupation or occupation area that typically requires education below the baccalaureate level. Skill competencies are defined as a concept, skill, or attitude that is essential to an occupation; the level of attainment or performance established for a skill competency is a skill standard. Because these terms tend to be used interchangeably in practice, the term “skill competencies” is used to refer to both skill competencies and skill standards.⁵

Wills⁶ emphasized that content standards state what learners should know and be able to do, whereas performance standards describe how well learners should know or be able to do something. Thus, state/program content standards address skills and knowledge, while instructional performance standards address levels of learning. In practice, there is no

² Boyer, Ernest. <http://www.ed.gov/offices/OVAE/CTE/2pgperk.html>

³ Lynch, Richard L., *New Directions for High School Career and Technical Education in the 21st Century*, ERIC Clearinghouse on Adult, Career and Vocational Education, Center for Education and Employment, Ohio State University, Columbus, OH

⁴ *Ibid* 1.

⁵ *Ibid*. 1.

⁶ Wills, J. *Standards: Making Them Useful and Workable for the Education Enterprise*. Washington, DC: Office of Vocational and Adult Education. US Department of Education, 1997. (ED 431 461)

“standard” standard; terminology is often inconsistent and both content and performance standards may separate or mix academic, employability and technical standards. Technical standards cover industry core standards, occupational “family” standards, and occupationally specific standards (typically as in a traditional CTE program).

Dr. Kenneth Gray⁷ has suggested that there are two programmatic goals of secondary CTE programs, to include

- 1) **Performance** goal: entry-level occupational competence, and
- 2) **Outcome** goal: transition to full-time employment

These goals have not changed significantly in the past 20 years, but what has changed is available technology and its effect on the nature of work, the organizational structure of school systems, and the development of instructional standards and curriculum modifications to reflect those standards. In addition, student performance expectations/criteria and the use of formal accountability systems documenting performance has become commonplace. Accountability emphasizes performance standards (how good is good enough) and student performance (how close are students to meeting the standard).

Accountability and Program Performance Systems

Scott Willis, Education Update⁸, stated in November 1999 that “Educators, schools, and districts are under constant pressure to show results that will convince policymakers and the public that they’re effective.” Forces influencing high school career and technical education have been described by Richard Lynch and include the new economy, public expectations, new cognitive science research about learning, and a variety of high school reform movements.⁹

Public expectations, broadly inclusive of parents, legislature, industry, and the community, held the widespread perception that U.S. schools are not nearly as good as they need to be—and therefore must be “held accountable.” As a result, in reauthorizing the Elementary and Secondary Education Act (No Child Left Behind Act), the federal government has required all states to establish accountability systems that include high-stakes testing. This has led states, local districts, textbook publishers, and other providers of content for education to begin work aligning standards, curriculum, and assessments.

It has also focused attention on the problems associated with reporting and tracking student performance and the use of powerful electronic systems to provide administrative support and performance analysis information. Examples of three Arizona school districts use of databases and electronic reporting systems are included in the section titled Criterion-Referenced Tests.

⁷ <http://www.ed.gov/offices/OVAE/HS/gray.doc>

⁸ Willis, Scott. Education Update, “The Accountability Question,” Association for Supervision and Curriculum Development, Vol. 41, Number 7, November 1999, 1.

⁹ Ibid.

These electronic systems operate in isolation of each other, as is true of most systems in Arizona and nationally, largely because they lack data-sharing compatibility and have been developed independently. Connecting programs and systems still need to be developed if states wish to somehow link local district electronic systems into a coordinated state assessment and accountability system. Some software has been developed that enhances this possibility.

For example, Pearson Education Technologies (formerly NCS Learn <http://www.PearsonEdTech.com>) has developed a new product called Concert.¹⁰ Concert is technology-based tool that provides student information, instruction and assessment, and business office applications. It uses a web portal to facilitate student, teacher and administrative collaboration/communication, and manages standards-driven content, resources and assessment, and student performance information. The developers have constructed Core Standards which are an aggregation of key states' standards. The intent is to allow alignment with standards across the country and to link standards and content, including content from both Pearson and other publishers.

Standards from 17 states are included in the system and new states are coming on-line regularly. Plans also include integrating NovaNET and SuccessMaker (acquired by Pearson Education) instruction as supplemental content linked to Core Standards, assessment, and reporting capabilities. Fletcher states that "what is unique is the aggregation of the pieces and the ways in which they work together...and the fact that they are on the Web, enabling access anytime, anywhere, and making updating and upgrading easily accessible...The missing- but planned for- piece is the content."¹¹

Several types of state accountability systems have been developed. These systems have many elements, but generally include components such as performance standards, student assessment, other indicators of performance such as graduation and dropout rates, incentives and rewards, and school or district sanctions. Accountability systems examine a dearth of information, but none so closely as student assessment results.

As a general practice, administration of student assessment for state accountability purposes may be centrally controlled (as in nearly all states, including Arizona) or locally controlled. In either case, students at various grade levels are required to take a mandated, standardized test or battery of tests to assess attainment of the state's

Standards. These tests are referenced as "high stakes testing" and have triggered a national debate about what role the test(s) should have in school accountability.

¹⁰ Fletcher, Geoffrey H., *Igniting the Internet Revolution: A new Category for Education Technology* T.H.E.Institute.2001. 1-2.

¹¹ Ibid. 8.

Rationale to Support or Reject High Stakes Testing

Those supporting high stakes testing do so primarily because they believe that the state content standards reflect the desired curriculum, that mandatory tests encourage improved instructional objectives and serve as a motivator for improved student and teacher performance, and that test results can provide clarity/direction to post-test learning activities. Their support includes the premise that everybody (public, school personnel, parents, community) should have the same expectations for all students and that waivers or special testing (because of unique local school or student circumstances) should not be available. Proponents for alternative assessment believe that alternative assessment strategies address inequitable and mitigating circumstances at particular school sites, and that they assess both interim and long-term student accomplishments more effectively.

Assessment appeals to policymakers because it is relatively inexpensive, can be externally mandated, implemented rapidly, and offers “visible” results. Student performance results have become a cornerstone for state accountability reporting. Parent surveys, in recent years, show some parent opposition to using tests to make high-stakes decisions and that parents are worried about the stress it may place on their children. However, a more recent poll indicated that approximately 63% of parents felt that standardized tests provide some benefits (Association of American Publishers (AAP)).¹²

Regardless of negative viewpoints on accountability, Asche¹³ suggests that a vocational education performance indicator system can have positive aspects including:

1. Locally developed indicators can provide opportunities for school-based improvement and the development of shared goals and values.
2. Indicators can be useful in monitoring policies and practice and improving schools.
3. Indicators offer an opportunity for vocational education to be included in educational reforms.

Those opposed to high-stakes testing primarily criticize performance standards that are set too high, content standards that are not aligned to the curriculum or the selected test, inconsistent methods of interpretation and test results analysis, and tests that do not accommodate second-language students and other special needs groups (i.e. low performing, low income, mobile populations, and handicapped).

Many commercial test developers and organizations specializing in assessment (e.g. Far West Laboratory and the Center for Research on Evaluation, Standards, and Student Testing) counter this criticism by providing alternative assessments, second language copies, out-of-level testing materials/guidelines and accommodation guidelines for test

¹² Olson, Lynn. (2000b) Test-makers' poll finds parents value testing. *Education Week*, 8(2), 16.

¹³ Asche, M. “Standards and Measures of Performance: Indicators of Quality for Virginia Vocational Education Programs.” Paper prepared for the teleconference “Preparing a Competent Work Force through Indicators of Quality for Vocational Education.” Blacksburg: Division of Vocational and Technical Education, Virginia Polytechnic Institute and State University, 1996. 6

administration. Out-of-level testing is one form of accommodation; it emphasizes testing students on content appropriate to their current level of functioning (which may be above or below their grade placement or age).

Other accommodation strategies are incorporated into several testing services materials. For example, The American College Testing Service's "Policy for Documentation to Support Requests for Testing Accommodations on the ACT Assessment"¹⁴ (<http://www.act.org/aap/disab/policy.html>) is an extensive application guide with teacher testing guidelines. Comparable resource materials are available from CTBS, Riverside (ITBS), Harcourt Education Measurement, and others. A copy of the ACT Testing Accommodations application guide is in Appendix A.

Beyond test service guidelines, some states have adopted special tests and test procedures to accommodate special needs students. For example, since 1989, South Carolina has used the Palmetto Achievement Challenge Tests for grades 10-12 basic skills assessment. In addition, this state has developed a portfolio-based assessment system (PACT-Alt) to meet the needs of students with *significant* disabilities who cannot participate in the regular assessment program—even with accommodations or modifications.

The portfolios address performance on the South Carolina Curriculum Standards for grades 3-8 in English, math, science and social studies. The system includes scoring rubrics with progress documentation strategies and instructor tutorials¹⁵ on how to use the rubrics. All materials are scored near the end of the school year (April-May). The SC Department of Education materials include an excellent alternative assessment portfolio resource guide.¹⁶ An on-line version is at <http://www.ihdi.uky.edu/mcrrc/> and a sample of the PACT-Alt Scoring Rubric is included in the Rubrics section of this report. With nominal modification, the rubric could serve as a special populations rubric model for Arizona.

The Mid-South Regional Resource Center at the University of Kentucky works closely with South Carolina and other member states that include New York, Rhode Island, Tennessee, Virginia, Washington, and Washington D. C. These states also are members in another group, the Inclusive Large Scale Standards and Assessment. ILSSA is a pool of professionals coordinated through the [Human Development Institute](http://www.ihdi.uky.edu) (<http://www.ihdi.uky.edu>) at the University of Kentucky and across the nation who partner with Measured Progress (Dover, NH)¹⁷ to form nationwide teams with expertise in working with students with disabilities.

The teams have extensive experience in developing inclusive large-scale assessment systems according to a state's particular need(s). The group works to assure that students with significant challenges are represented in the accountability system and the state assessment system reflects the most current research on accommodations, alternate assessment and reporting practices.

¹⁴ ACT website: <http://www.act.org/aap/disab/policy.html>

¹⁵ *South Carolina Palmetto Achievement Challenge Tests Alternate Assessment Portfolio Guide*, South Carolina State Department of Education. <http://www.sde.state.sc.us>

¹⁶ *Ibid.*

¹⁷ Mid-South Regional Resource Center <http://www.ihdi.uky.edu/mcrrc/>

ILSSA encourages professionals to develop testing models and guidelines for low-achieving, disadvantaged, handicapped, and other special needs populations to assure fair testing practices. Many authors agree that the area of special needs testing is poorly developed, selectively punishes low-performing students, does not reward incremental improvements, and fails to foster an intrinsic interest in the subject matter. Improved performance, they argue, will result when such barriers are removed; the primary goal of ILSSA is to reduce barriers.

Olson¹⁸ concluded that: "The risk of undermining the future of students with limited English proficiency is significant and...until American education becomes more equitable, high-stakes testing will continue to show massive bias and differential outcomes." Mac Iver¹⁹ states that "traditional evaluation systems often do not adequately recognize the progress that educationally disadvantaged students make, because even dramatic progress may still leave them near the bottom of the class in comparative terms or far from the 'percent correct' standard needed for a good grade." To counter these criticisms, Baltimore Public Schools' "Incentives for Improvement" program uses an incentive system to encourage individualized, doable, short-range learner goals and provides certificates and other awards for improvement, thus encouraging student successes and recognizing interim progress.

High-stakes test critics cite test modes that emphasize reliance on memorization of facts, for emphasizing low-level skills and piecemeal knowledge, and for being biased in favor of white, middle-class children.²⁰ Elmore, Abelman, and Furhman (1996) concluded that schools can be held accountable only for those factors they can control, but should not be held accountable for student background or prior achievement (which) institutionalizes low expectations for poor, minority, and low-achieving students.²¹

Critics further decry the negative impact low scores have on the public's perception of school effectiveness. Some argue that high-stakes tests should be combined with other periodic tests to more fully measure the total curriculum and student outcomes. Further, they are concerned that teachers will only "teach to the test," and, thus, narrow the curriculum. Teacher organizations, particularly the American Federation of Teachers (AFT) are concerned that teachers lack time and resources to learn effective teaching strategies to apply to standards that must be taught, and tested.²²

Arguments about who is responsible for student achievement have emerged. Some believe that student achievement should be everyone's responsibility and that accountability should be spread throughout the organization and not rest *solely* with the instructional staff. They are concerned that accountability limits teacher control over curriculum and instruction and suggest it reduces the quality of their professional lives. Opponents also argue that test scores and student outcomes are *not automatic indicators* of teaching effectiveness and that

¹⁸ Olson, Lynn. (2000a). High-stakes tests jeopardizing Hispanics, panel warns. *Education Week*, 7(12), 7.

¹⁹ Mac Iver, quoted in ERIC *Education Reforms and Students at Risk: A Review of the Current State of the Art* January 1994. 1

²⁰ *Ibid*

²¹ Elmore, R. F., Abelman, C. H., & Furhman, S. H. (1996) The new accountability in state education reform: From process to performance. In H. F. Ladd (ed.) *Holding Schools accountable: Performance-based reform in education* (pp.93-94). Washington, DC: The Brookings Institution.

²² Bradley, Ann. (2000, July 12). Union heads standards warnings. *Education Week*, 7(12), 1,20-21.

effectiveness should be measured with other performance criteria and not just with student achievement scores.

Finally, assessment-based accountability models currently in use are suspect in terms of whether or not they really show improvements in education. Critics point out that while improved performance on these measures does increase over time the results are suspect because improvements may be:

1. linked to the use of old norms,
2. the repeated use of test forms year after year,
3. the exclusion of students from participating in accountability testing programs, and
4. the narrow focusing of instruction on the skills and question types used on the test.²³

State Accountability Reporting

Model 1: School-based, aggregated data

States such as Virginia, Kentucky, Maryland, Arizona, Michigan, Colorado, Missouri, and others have content standards and standards-based core assessment systems. Performance standards to specify desirable attainment levels are components of these systems.

Wonacott²⁴ reports that many states have adopted industry-based skill standards as part of their core assessment systems. In some cases, industry skill standards are imbedded as in the case of California's Career-Technical Assessment Program (C-TAP); Ohio's Integrated Technical and Academic Competencies (ITACS); and Oregon's Certificate of Initial Mastery (CIM) and Certificate of Advanced Mastery (CAM). The mastery certificate is aligned with the Performance-Based Admissions Standards (PASS) for Oregon State's university system.

Many states, including Arizona, Virginia, Michigan and others, have implemented state content and performance standards. These states use mandated, state-administered test scores (usually based on standardized tests) as a component of their state accountability and accreditation system. Assessments are, theoretically, aligned to the state academic standards. Alignment refers to the degree of match between test content and the subject area content identified through state academic standards; it includes content match and depth match. Webb²⁵ states that "depth match refers to cognitive complexity prescribed by the standards and the cognitive complexity required by the assessment item/task,"

Lack of match between content and test items is the primary criticism against standardized test scores being used for accountability purposes. The correlation between attaining state standards and the level of student achievement in meeting performance standards is primarily the guideline used to determine school ratings. This is the case in slightly more

²³ Linn, Robert L., (2001) *Assessments and Accountability (Condensed version)* Practical Assessment, Research & Evaluation, 7(11).3 ERIC Clearinghouse on Assessment and Evaluation ISSN 1531-7714.
<http://ericae.net/pare/getvn.asp?v=7&n=11>.

²⁴ Wonacott, Michael E., *Standards: An Embarrassment of Riches In Brief: Fast Facts for Policy and Practice* National Dissemination Center for Career & Technical Education Washington: DC 2000 (4) I

²⁵ Webb, N. L. (1997 and 1999). *Research Monograph No. 6: Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education and Alignment of Science and Mathematics Standards and Assessments in Four States*. Washington, DC: Council of Chief State School Officers

than 50% of the states. Typically, low ratings become the basis for school sanctions that, in turn, often generate negative public perceptions of school efficiency and effectiveness.

Critics argue that unless there is a match between content and test items the test is invalid initially and, therefore, the results should not have such impact. La Marca²⁶ summarized that “The benefits of confidence, fairness, and defensibility to students and schools outweigh the costs (of alignment).” A copy of his article titled *Alignment of Standards And Assessments as an Accountability Criterion* is included in Appendix F. The article summarizes alignment concepts, methodologies, and issues.

There is not agreement about how performance data is used for State accountability purposes. Critics believe that performance improvement often is unnoticed and/or unrewarded in accountability systems. They argue that limited use has been made of cash incentives and recognition programs for schools showing improvements in performance. It is important to note that many state-run accountability systems do not have a mechanism to reward “selective” good performance. For example, a generally low-performing school may be making significant progress with disadvantaged students but overall not perform well on a standardized test, resulting in the school being classified as “not meeting state standards.”

Another criticism is that many accountability systems do not recognize disaggregated (individual) performance data but review, instead, composite test scores—usually by grade level and content area. Critics believe that recognition, incentives, and awards should be given for improvement in disaggregated categories as well as for overall, satisfactory performance. Kentucky legislation acknowledged that some schools deal with high numbers of disadvantaged student populations and modified its legislation to address this issue.²⁷ Their approach recognizes interim improvement, while simultaneously requiring schools with low performing students to move more aggressively toward attaining the State standard, which states that:

1. All schools must reach 100 on a 140-point scale by 2014;
2. All schools must make steady progress to attain the scale in the interim;
3. Low-scoring schools must make *more progress* than other schools during the same period.

Model 2: School-based, disaggregated data

An alternative model that utilizes disaggregated data as an accountability model is provided through the Commission on Accreditation and School Improvement, commonly known as the “North Central.” This national organization uses a peer review process to accredit K-12 schools. Their program, called Transitions, is a school improvement model based on individual student performance, not composite scores based on a standardized test. It is intended to ensure that every student is prepared for successful transition to the next school or life transition. Finally, it represents the “exemplary level” of the NCA Performance Accreditation Framework.

²⁶ La Marca, Paul M. (2001) *Alignment of Standards and Assessments as an Accountability Criterion* ERIC Clearinghouse on Assessment and Evaluation, ISSN 1531-7714. 6

²⁷ Willis, Scott. *The Accountability Question* Education Update ASCD 41(7) Alexandria, VA. November 1999. 5

Unique characteristics of the program include:

- **Requires specific improvements** in student performance;
- **Credentials students** by validating each student's present performance in academics, employability skills, and career awareness/exploration;
- **Requires an articulation plan** describing how sending/receiving schools will communicate and collaborate to improve the performance of individual students.²⁸

The program requires each school to develop a written 5-year School Improvement plan, develop *individual* student intervention plans, align/restructure site support systems, and select processes to support Transitions goals. A total of 142 pilot schools in 13 states of a 19-state region are working to develop individual student rubrics for instructor evaluation of progress in each of five areas.²⁹

Credentialing areas for middle and high schools in the Transitions program are comparable to many of the CTE cross-program competencies and/or state Academic Standards and Workplace Skills discussed in a later portion of this report. NCA credentialing areas include:

1. Reading Comprehension
2. Writing
3. Mathematics
4. Reasoning, Thinking, or Information Processing Skills
5. Employability Skills
6. Career Awareness

²⁸ North Central Association, <http://www.ncacsi.org/transitions/>

²⁹ Ibid.

Model 3: Transition or Gap Year Provisions

Another model, adopted by the Rochester, New York School Board, is the “Pathways” plan. Students have the option to finish school in three years, or they may remain in high school for up to five years if they need more time to master state-mandated standards. Transition or gap year provisions are common practice in private schools and many schools in Europe.³⁰ This model encourages students to stay in school, have improved opportunity to excel, and reduce student hopelessness when/if they are unable to meet state guidelines. The extra year in school provides students with additional study and preparation time to pass state standards.

In other countries that have national standards exams, students know that their futures depend on their test scores. Many foreign schools provide different types and levels of competency certification to better inform employers of student strengths. Factors such as length of time in school, work experience, and a work record of competence and reliability encourages potential employers to more readily accept graduates.

Clearly, there is no standard practice among US and other countries. Robertson (2000)³¹ has suggested seven items for policymakers to consider to enhance the validity, credibility and positive impact of assessment and accountability systems while “minimizing their negative effects” that include:

1. Provide safeguards against selective exclusion of students from assessments.
2. Make the case that high-stakes accountability requires new high-quality assessments each year that are equated to those of previous years.
3. Don't put all of the weight on a single test. Instead, seek multiple indicators. The choice of construct matters and the use of multiple indicators increases the validity of inferences based upon observed gains in achievement.
4. Place more emphasis on comparisons of performance from year to year than from school to school. This allows for differences in starting points while maintaining an expectation of improvement for all.
5. Consider both value added and status in the system. Value added provides schools that start out far from the mark a reasonable chance to show improvement while status guards against institutionalizing low expectations for those same students and schools.
6. Recognize, evaluate, and report the degree of uncertainty in the reported results.
7. Put in place a system for evaluating both the intended positive effects and the more likely unintended negative effects of the system.

Performance Indicators for CTE Program Accountability

Because of federal funding requirements, Arizona Career and Technical Education (CTE) programs, funded through the Carl D. Perkins 1998 Act (Perkins III), are required to establish a system of standards and measures to assess vocational education programs and report attainment on state performance measures. States must also provide an annual report to the US Department of Education, Office of Vocational and Adult Education (OVAE), comparing

³⁰ Robertson, Anne S. “High-Stakes” Testing: New Guidelines Help Direct School Change. NPIN Parent News for November-December 2000. 2 <http://npin.org/pnews/2000/pnewl00/intl00b.html>

³¹ Ibid. 4

the status of occupational programs with the goals identified in the state's annual performance plan.

Federal funds used to support local district programs have mandatory accountability criteria (Perkins, Section 113(b)(2)(A)). The criteria address program evaluation, rather than student performance per se. Student performance aggregated school assessment data is used as an accountability indicator (measure 1.1 and 1.3 below) for state/federal reporting purposes. The performance measures are reviewed annually and performance targets are revised upward for the ensuing year, per federal guidelines. The present Arizona performance measures and indicators (2003) are shown in Table I.

Table I: Core Indicators and 2003 Performance Measures

Indicator 1.	<i>Student attainment of challenging state-established academic, and vocational and technical skill proficiencies [Sec. 113 (b)(2)(A)(i)].</i>
Performance Measure 1.1	20.00 % of CTE program concentrators who leave secondary education in the reporting year will meet or exceed all the state standards as assessed by the AIMS test.
Performance Measure 1.3	55.00% of CTE program concentrators who leave secondary education in the reporting year will (1) pass the state-adopted proficiency assessment OR in the absence of a state proficiency assessment (2) pass at least 80% of the total program competencies and are documented as attaining at least 80% of the occupational Level III program competencies in an approved CTE program
Indicator 2.	<i>Student attainment of a secondary schools diploma or its recognized equivalent, proficiency credential in conjunction with a secondary school diploma, or a post-secondary degree or credential [Sec. 113 (b)(2)(A)(ii)].</i>
Performance Measure 2.1	91.00% of CTE program concentrators will leave high school due to graduation in the reporting year.
Indicator 3.	<i>Placement in, retention in, and completion of, post-secondary education or advanced training, placement in military service, or placement or retention in employment [Sec. 113 (b)(2)(A)(iii)].</i>
Performance Measure 3.1	41.56% of CTE program completers who graduated in the previous year were placed in post-secondary education or advanced training, military service or employment.

Indicator 4.	<i>Student participation in and completion of vocational and technical education programs that lead to nontraditional training and employment [Sec.113 (b)(2)(A)(iv)].</i>
Performance Measures 4.1 & 4.2	31.13% of enrollment in nontraditional CTE programs will be nontraditional genders.
Performance Measures 4.3 & 4.4	26.93% of completers in nontraditional CTE programs will be nontraditional genders.

The National Center for Education Statistics, February 2000,³² states that CTE performance accountability systems are intended to:

- Include four core indicators that **measure student performance and post-vocational education experiences** in further education, training, and employment;
- **Set performance levels** for the four vocational outcomes, including student attainment of skill proficiencies; and
- **Measure and report the performance** of the states on the indicators.

Measuring student performance and validating attainment of skill competencies for performance measure reporting is accomplished through site visits. School site audits are conducted annually to address performance measures, including reviewing student performance records, support services documentation (i.e. IVEP records), and graduation/placement information. Data from these audits are the basis of the ADE report to the USOE, OVAE for its annual accountability report and for Department of Education program approvals for the ensuing funding cycle.

For detailed historical information and a discussion of assessment and accountability reporting issues related to the Perkins III core indicators, readers may wish to access an excellent paper written by David Stevens at Ohio State University.³³

General Purposes of Student Assessment

“Accountability is based on the teacher’s adjusting practice to maximize the likelihood of student success and to minimize student failure.”³⁴

³² US Department of Education, National Center for Education Statistics E.D.Tab., “Occupational Programs and the Use of Skill Competencies at the Secondary and Postsecondary Levels, 1999, NCES 2000-023, by Basmat Parsad and Elizabeth Farris. Bernie Greene, project officer. Washington, DC: February 2000, 1.

³³ Stevens, David W. (2001). *21st Century Accountability: Perkins III and WIA Information Paper 1002*. National Dissemination Center for Career Technical Education. Ohio State University. Columbus OH.
<http://www.nccte.org/publications/secure/index.asp#21stCentury>

³⁴ Willis, Scott. *The Accountability Question Education Update* ASCD 41(7) Alexandria, VA. November 1999. 8

Student assessment is an accountability indicator in the CTE performance measures. However, student assessment viewed from the perspective of the learning environment has somewhat different purposes than when used as a state accountability measure. The New York State Education Department published a statement of general purposes of assessment.³⁵ Included were:

- **To Plan Instruction** – If achievement is assessed before instruction, instruction can be tailored to meet the needs of students. In addition, the students will better understand the specific objectives for instruction.
- **To Motivate Students** – Most students will exert a greater effort to learn if they know how their achievement will be measured.
- **To Evaluate Instruction** – The extent to which students attain an objective is one indication of the effectiveness of instruction.
- **To Assist Learning** – Some assessment techniques provide opportunities for students to apply what they have learned, thereby reinforcing instruction.
- **To Measure Achievement** – Perhaps the most obvious reason for measuring achievement is to assign grades which are fair and accurate measures of student growth.

The National Forum on Assessment, in *Principles and Indicators for Student Assessment Systems*³⁶ developed seven ethical premises to reinforce the general idea that evaluation should serve the cause of appropriate student instruction. Principle 1 states that the “*primary purpose of assessment is to improve student learning.*” Such assessment may include classroom-based and large-scale assessment activities.

Seventeen indicators of this principle include:

1. Assessments are based on curriculum and desired learning outcomes that are clearly understood by students, educators, and parents.
2. Assessment practices are compatible with current knowledge about how learning takes place and allow for variety in how students learn.
3. Assessment systems enable a process of continuous feedback for the student.
4. Most assessments allow students to demonstrate understanding by thoughtfully applying knowledge and constructing responses.
5. Assessment systems allow students multiple ways to demonstrate their learning.
6. Assessment systems include opportunities for individual and group work.
7. Classroom assessments are integrated with curriculum and instruction.
8. Teachers employ a variety of assessment methods and obtain multiple forms of evidence about student learning for planning and implementing instruction and for evaluating, working with, and making decisions about students.
9. Teachers can explain how their assessment practices and instruments help improve teaching and how they provide useful information for working with students.
10. Student's self-reflection and evaluation are part of the assessment system.
11. Schools establish procedures for enabling classroom-based student assessment information to follow each student from year to year.
12. Assessment methods, samples of assessments, scoring guides or rubrics, and examples of work of varying kinds and quality are discussed and understood by students.

³⁵ *Assessing Achievement in Home Economics Education*, New York State Education Department, Albany, NY 1991.

³⁶ *The National Forum on Assessment: Principles and Indicators for Student Assessment Systems*. National Center for Fair and Open Testing (Fair Test), Cambridge, MA, 1995.

13. Scoring guides (rubrics) state in positive terms what students can do and enable users to analyze student strengths and needs in order to plan further instruction.
14. Educators make clear to students the uses and consequences of each assessment.
15. Teachers use current principles and technical concepts of assessment, particularly validity and reliability, in developing and analyzing their classroom assessments.
16. Multiple-choice and short-answer methods are a limited part, in time or impact, of the total assessment system.
17. Assessments intended to rank order students or compare students with each other are not a significant part, in time or impact, of the total assessment system.

Most authors agree that when properly developed and administered, student assessments can:

- Reinforce instruction
- Stimulate student interest
- Gauge student readiness for new learning
- Measure achievement
- Indicate effective instruction

Behuniak³⁷ summarizes the differences in assessment purposes by stating “This apparent paradox might appear to be hopeless until one realizes how many successful applications of educational assessments occur every day, despite the complications caused by these cross purposes. Every capable teacher can provide numerous examples of ways in which formal and informal means of assessing student achievement helped to diagnose a learning problem, document progress, identify an effective instructional approach, and produce numerous other desirable outcomes...(and are) enhancements to the educational experience.”

The issue of *how to measure* student performance centers on the availability and acceptance of a variety of performance measures, including state-mandated performance assessment(s) in either/both academic, workplace, and vocational competencies. Use of test scores as an accountability measure appears, on the surface at least, to be at cross-purposes with the instructional purposes of assessment. Conversely, given that one of the purposes of assessment is to measure achievement, such measures could be and are used to aggregate student data for use in both instructional and accountability reporting. Clearly, though, there is not agreement on this double use of student performance data.

Assessment Principles

1. Assessment is inherently a process of professional judgment (McMillan).³⁸

The foundation principle of the assessment process involves realizing that teachers use professional judgment to make professional interpretations and decisions about the measurement of student performance. This viewpoint supports the precept that what's tested

³⁷ Behuniak, Peter. (2002). *Phi Delta Kappan Consumer-Referenced Testing*. PDK 84/3. Bloomington, IN. 201.

³⁸ McMillan, James H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical assessment, Research & Evaluation*, 7(8). <http://ericae.net/pare/getvn.asp>?

and what's expected matters much more than which way the testing is done. Practices such as machine scoring and use of multiple-choice test items seems "objective," but even these forms are based upon the professional expertise and judgment of the teacher who selects both content and test structure.

Whether that judgment occurs in constructing test questions, scoring essays, creating rubrics, grading participation, combining scores, or interpreting standardized test scores, the essence of the process is making professional interpretations and decisions. Understanding this principle helps teachers and administrators realize the importance of their own judgments and those of others in evaluating the quality of assessment and the meaning of the results (interpretation of the description or scores).³⁹

2. Assessment is a necessary and fundamental component of the teaching process.

Assessments audit what students have learned and can motivate students. If constructed as authentic measurements assessments can provide feedback and opportunities to improve curriculum, instruction and/or learning. Assessment is an integrated experience within the instructional program.

This viewpoint supports the precept that test scores should be used to improve the instructional program. Regardless of the assessment type used, students should understand what will be assessed, how the assessment will be measured, and how the performance results will be used. Others (i.e. teachers, parents, etc.) need to understand how "testing" relates to any one of several purposes (i.e. accountability, student strengths/weaknesses, school/district community reporting, etc.).

3. Assessment in its broadest meaning is different than evaluation.

Assessment is different than evaluation. Assessment implies "taking a look" at something over an extended period of time, rather than "evaluating" something at a specific point in time – as is the case with many state-wide (high-stakes) accountability measures which interpret composite scores for accountability purposes.

Assessment emphasizes individual student achievement, not school or program evaluation and is considered part of a continuous process. It links the teaching/learning process to desired learner outcomes. Shepard (2000)⁴⁰ concluded that "...considerable recent literature...has promoted assessment as something that is integrated with instruction, and not an activity that merely audits learning." Thus, assessment is usually an intermittent, integrated process, rather than a mandated, single purpose effort.

Authentic assessments (alternative strategies vs. traditional standardized objective tests) such as portfolios, oral and written presentations, and scenarios requiring problem-solving skills are evaluative strategies used widely in vocational/technical programs. Industry credentialing and external (third-party) testing services are another option for the CTE

³⁹ Ibid.

⁴⁰ Shepard, L. A. (2000). The Role of Assessment in a Learning Culture. Paper presented at the Annual Meeting of the American Educational Research Association. Available <http://www.aera.net/meeting/am2000/wrap/praddr01.html>

programs. Many industry certificates exist, but are not found in one definitive listing. Additionally, there is not consensus on whether they should be for all programs and no common format exists for such materials. Both authentic assessment and industry certification is discussed in more detail at a later point.

Standardized tests (which may or may not include a performance assessment component) are another dimension of assessment administered as a “high-stakes” (single purpose) test. Conversely, a single-purpose standardized test may be one part of the continuous testing process and, in that case, test results are coupled with other assessments to determine student achievement. Behuniak⁴¹ references the latter strategy as a “multi-tier” assessment approach in which different tests serve different purposes, ranging from accountability to instructional purposes. He takes issue with single-purpose tests and argues that students and teachers are not prepared and suggests “It is illogical and counterproductive to implement high-stakes assessments before teachers have had reasonable opportunity to become familiar with the covered content and introduced appropriate instruction in the classroom.”⁴²

4. Assessment decision-making is influenced by a series of tensions.

McMillan⁴³ suggests that the purposes, uses, and pressures are conflicting and result in tension for teachers and administrators. Typical tension areas include:

- Student performance vs. mandated large-scale testing
- Learning vs. auditing
- Formative (informal and ongoing) vs. Summative (formal and at the end)
- Criterion-referenced vs. norm-referenced
- Value-added vs. absolute standards
- Traditional vs. alternative
- Authentic vs. contrived
- Speeded tests vs. power tests
- Standardized tests vs. classroom tests

5. Certification and high-stakes testing programs are misnomers for “assessment.”

Year-end or semester-end single tests may occur as one part of the assessment continuum. When presented as year-end or semester-end assessments, industry certification tests and state-mandated tests are administered at a time in the school year that provides little, if any,

⁴¹ Behuniak, Peter. (2002). Phi Delta Kappan *Consumer-Referenced Testing*. PDK 84/3. Bloomington, IN. 202.

⁴² Ibid. 206

⁴³ McMillan, .2.

time for modification of the instructional program and/or corrective action on the part of the student.

These test administrations are sometimes called assessments, but, in fact, are evaluations when scores are interpreted only for accountability purposes. Industry-developed tests have “evaluation” purposes leading to certification/non-certification of student performance in an industry-specific skill area. Such tests document skill attainment (end-results) and generally do not document other “soft skills” and attitudes schools emphasize as part of the instructional program.

Critics point out that the test provide little, if any, opportunity for creative expression, teaming, and demonstration of other personal attributes and leadership skills that potential employers state they want in new employees and that many states include in their academic/vocational standards. These tests also function as school and/or program evaluation data when used for accountability purposes to indicate composite “pass rates” for respective programs.

Standardized Tests

Obviously, priorities need to be made and trade-offs are sometimes necessary in making assessment selection choices. The Phi Delta Kappan magazine recently featured special sections on standards and testing. In it, Meier⁴⁴ states

That a standardized one-size-fits-all test could be invented and imposed by the state, that teachers could unashamedly teach to such a test, that all students could theoretically succeed at this test, and that it could be true to any form of serious intellectual or technical psychometric standards is just plain impossible. And the idea that such an instrument should define our necessarily varied and at times conflicting definitions of being well educated is—worse still—undesirable.

There is not agreement among professionals on the value of standardized or end-term tests. However, they are one method of determining whether teachers and students are doing their jobs. Curriculum and test alignment activities encompass adjusting the curriculum to reflect state/district standards that, in turn, students must learn in order to pass the test instrument(s). Both standardized tests (AIMS, Stanford 9) and locally-developed tests (usually criterion-referenced) are used in Arizona CTE programs. The advantages/disadvantages of each test type and criteria for selection of each follows.

Standardized tests:

Standardized tests are used to provide “comparable” test results related to specific groups and/or content. Test administration and scoring procedures are established by the commercial test developer who also provides scoring services on a fee basis. The tests are secure tests with the same form(s) used nationwide by any school/district contracting for the test administration.

⁴⁴ Meier, Deborah. (November 2002) *Standardization vs. Standards*. Phi Delta Kappan 84(3). Bloomington, IN 192.

Norms relating the results to pre-established groups or content are available for interpretation/analysis purposes. Some companies have or have in process alternative test forms for special populations; some companies also have second language test forms available.

PRO

1. Standardized tests, in general, predict less well than grades, but standardized tests plus grades predict much better than grades alone, because grades vary considerably from teacher to teacher.
2. Standardized tests may be used to help establish homogeneous groups (appropriate class placement)
3. Standardized tests have an advantage of high quality of items and careful planning of content.
4. Sample testing from a large bank of questions (mostly multiple choice), as in national assessments, provides a way of comparing student performance to a reference group (regional/national or special purpose).

CON

1. Specific tasks may require more test items than are available in existing test item banks and/or standardized tests.
2. Standards of satisfactory achievement are not reflected in norming tables. Norming tables reflect the performance/achievement of populations geographically, not standards attainment levels.
3. Norm groups or referent groups often do not match the group tested.
4. Standardized tests encourage homogeneous rather than heterogeneous groupings (potential abuse of results).
5. "Teaching to the test" (coaching) may occur (and limits the curriculum).
6. Many students lack test taking skills and/or "freeze" when confronted with a standardized, mandatory test.
7. Issues of culture, language, ethnicity, and equity are inherent criticisms of many standardized tests.
8. Some standardized tests lack forms/procedures to accommodate special populations such as second language, physically challenged, and/or learning disabled students.
9. May be too simplistic and not adequately measure a student's ability to think and solve problems. (New York Times)

Criteria for Standardized Tests:

NOCTI has suggested four criteria for standardized tests, indicating tests should:

- Include **measures of both technical knowledge and skills** and technical literacy aligned with state and national standards.
- Be designed to **measure student progress** against clear and rigorous technical and technical literacy standards.
- **Meet criteria for quality assessments** (such as the American Education Research Association, American Psychological Association, and National Council on Measurement in Education Standards for Educational and Psychological Testing). Tests must be:
 - Valid
 - Reliable
 - Fair and non-biased
 - Secure
- **Be benchmarked** at the national, regional, state, and/or local levels.⁴⁵

McMillan⁴⁶ summarized suggestions and guidelines from fifteen recent research sources. He found agreement among the authors that:

Good assessments:

- enhance instruction,
- are valid,
- are fair and ethical,
- use multiple methods,
- are efficient and feasible, and
- appropriately incorporate technology.

⁴⁵ NOCTI, *Using Standardized Test Data To Improve Instruction In Career-Technical Education, A Perspective for Practitioners*. (undated), 4.

⁴⁶ McMillan, James H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical assessment, Research & Evaluation*, 7(8). <http://ericae.net/pare/getvn.asp>? 4-5.

Criterion-Referenced Tests

Nationwide, consortiums and school districts have developed sponsored tests for both academic and vocational programs. For the most part, these tests are identified as “criterion-referenced tests” and are unique to the using district or consortium. It is possible to develop criterion-referenced test item banks that use multiple-choice, matching, performance, and scenario test item construction strategies discussed in a later portion of this report. The importance of the test item selection rests in the items’ ability to measure mastery of the specified instructional content.

Criterion-referenced tests are most commonly integrated with instruction. They are introduced before, during and after completion of each instructional unit to check students’ prerequisite skills, diagnose possible learning difficulties, prescribe the needs of possible subsequent instruction and/or determine mastery. To be effective, instructional objectives must be specified; test items must be written to be consistent with the learning objectives and must parallel the instructional content. Criterion-referenced tests contain items referenced *primarily* to cognitive domains, as compared to performance-based tests referenced to both cognitive and psychomotor domains.

PRO

1. Criterion referenced tests can be constructed to meet specific instructional goals. Enhances instruction.
2. Accommodates unique course/class content. Tailored to each school site.
3. Criterion tests may test for single tasks, multiple tasks, or multiple competencies.
4. Criterion-referenced tests respond to accountability issues.
5. Test items are more likely to be relevant and reflect the curriculum. Test questions match significant content to be learned.
6. Criterion-referenced tests can be shown to have predictive validity to a large extent. A criterion-referenced test cannot guarantee that students can perform at a specific level of competence on these instruments.
7. Contains elements that are valid, fair and ethical.
8. Use multiple methods.
9. Efficient and feasible to administer and appropriately incorporates technology.

CON

1. Requires constant review/revision to remain current with curriculum.
2. The criterion level of expected performance is difficult/time consuming to determine. Items are time consuming to construct.

3. The identification of subject matter (content) and appropriate levels of cognitive ability must be completed before test items can be developed.
4. Before test items can be developed, the instructional objectives must be developed and agreed upon by all instructors.
5. Often not applicable across variety of same-subject teachers.
6. Planning time is required to assure that test specifications are developed and the number of test items selected for the test reflect the relative importance of curriculum areas being taught.
7. Testing the reliability of the assessment instrument is a complicated process. It includes identifying the number of items needed to yield reliable evaluation of each of the specific instructional objectives covered by the list, and the proportion of items to accurately measure the level of mastery. (It represents precision of measurement). Reliability is no guarantee of validity.
8. May not meet reliability/validity tests. Testing the validity of the assessment (delineates what a test measures and how well it measures what is being tested) necessitates item analysis and review of cognitive domains being measured. (Face validity - the surface appearance of validity - is not enough.) To be valid, a test must be reasonably reliable.
9. Item analysis includes scrutiny of the following factors:
 - a. Item difficulty (percentage passing and appropriate difficulty)
 - b. Item validity (Relationship between item response and criterion performance and having high discriminating power)

Criteria for criterion-referenced tests:

1. Determine if items are for single tasks, multiple tasks, or multiple competencies.
2. Delimit the specific instructional goals addressed by the item(s)
3. Develop/select test questions that match significant content to be learned.
4. Assure that test items are valid, fair and ethical.
5. Construct items/select tests that use multiple methods.
6. Verify that test administration and scoring practices assure consistency.

Uses of criterion-references tests differ from district to district in Arizona and nationally. For example, Tucson and Deer Valley Districts have criterion-referenced district-wide exit exams for several academic areas, primarily for elementary grades. None of the districts surveyed have district-developed, criterion-referenced on-line testing capability. Several of 14 districts surveyed are using V-TECS item banks, NOCTI, and/or departmental tests of some type, but for the most part this is not typical of Arizona CTE programs. Conversely, Arkansas requires schools to first look for assessments using a nationally recognized certification provider (i.e. ASE for automotives). If one is not available, the appropriate NOCTI test *must* be used. If

neither is available, then the school must develop their own assessment first using V-TECS items (if available) or district-developed items.

V-TECS maintains a large test item bank that can be accessed by participating schools/districts. The test item bank uses multiple-choice, matching, performance, and scenario items. One shortcoming of the V-TECS item bank is that reliability/validity study of test items in the bank has not yet been accomplished and that it is an unsecured test site.

California presently utilizes state-developed Assessments in Career Education (ACE). The assessment emphasizes honoring student achievement, not school or program evaluation, and uses two 45-minute sessions in which students complete multiple-choice, written-response questions or problem-solving tasks. There are no costs to students or schools for participating in the ACE – it is essentially a voluntary program to recognize individual achievement. Students are awarded special certificates of achievement by the California Department of Education.

Georgia, Maryland, Tennessee, and Virginia are developing end-of-course exams; other states are considering this option. Twenty-five states have differentiated diplomas; some of them specifically include career technical education. At present Florida, Georgia, Massachusetts, Ohio, Virginia, and West Virginia award differentiated diplomas.⁴⁷ Massachusetts is currently developing Certificates of Occupational Proficiency for 42 occupations that are the focus of high school CTE programs. Some other states use tests such as Work Keys to assess work-readiness of students. A more limited number of states incorporate AP or IB achievements as part of their assessment and diploma-granting strategies.

Promising Practices: Several school districts in Arizona have a history of using district-developed tests, including objective, criterion-referenced and performance types. For example, Tempe Union High School District, Apache Junction Unified District, and Glendale Union High School District have extensive materials, processes, and reporting mechanisms in place to support their student competency attainment, testing, and instructional program improvement efforts.

Example 1: Tempe Unified School District

Tempe Unified District has provided extensive professional development activities to assist teachers in developing competency-based instruction. Teachers utilize competency checklists, rubrics, and assessments (at the discretion of each teacher/program). The district maintains an on-line competency attainment report for each student. The site is a secure site in that once one teacher has certified competency attainment, another teacher is unable to change that record. The electronic system assumes that if the student has attained a competency, there is no reason to allow a “pass” to be changed to a “not attained” rating by another teacher. This places the burden of competency attainment verification on each teacher who, in turn, knows that her/his judgment directly affects the student’s performance and attainment in the next program sequence class.

⁴⁷ Wills, Joan. *Promoting New Seals of Endorsements in Career Technical Education* The National Association of State Directors of Career Technical Education Consortium, Washington, D.C. 2002. 8.

Tempe's data management system generates graduate transcripts and competency attainment reports for each student in each program for the District to use for student performance and state accountability reporting. Samples of their student accountability reports are included in Appendix B.

Example 2: Apache Junction Unified District

Apache Junction Unified District has provided extensive professional development for staff to implement standards, write instructional objectives, develop instructor test-item writing skills, and utilize on-line assessment systems. The District purchased an academic test item bank, setup, and hardware from McGraw Hill. Teachers are able to import and export test items to create AIMS-type assessments. The electronic system scans test results and prepares a variety of administrative/teacher reports.

Example 3: Glendale Union High School District

Glendale Union High School District utilizes criterion-referenced tests and provides teachers with multiple-choice item writing guidelines, shown later in this report. The District has provided extensive professional development for instructional staff who utilize district-developed objective and performance-based assessments for all vocational programs. Some tests are on secured sites; others are non-secured. Performance-based assessments include task descriptions, mandatory components, student directions, checklists, rubrics, and activity worksheets. Student assessments are scored by each teacher who then converts the performance rubrics, coupled with other class performance information, into a student grade for the particular class.

Each summer, a district rubrics workshop is held for district scorers who review a 30% sample of each district teacher's student performance-based assessments. Two trained scorers rate each sample student's assessment. These are "blind" ratings in which the rater does not know the school site or student name. Rater scores are analyzed for inter-rater reliability against the district standard of .80 co-efficient.

Scores are entered on a database that assembles composite scores for the district and generates a series of reports for competency reporting and analysis. Subsequently, modifications of the performance-based assessment instrument are made, if necessary, and the instructional program objectives are re-validated. The process is a dynamic process that changes if/when program standards, competencies, or instructional content changes.

Samples of two Glendale District non-secure performance-based assessment booklets (Multimedia Computer Applications and Child Development) are in Appendix C. The District has performance-based assessments for drafting, cabinetmaking, introduction to technology, business computer applications, multimedia computer applications, keyboarding, child development and family living.

Database capabilities/management: In the case of Tempe, Apache Junction, and Glendale Districts, administrative and database management systems are highly supportive, professional development release time and monies to support extended contracts is

available, and assessment system refinement activities are ongoing. All three districts maintain information system on-site personnel to support their assessment and documentation systems.

These districts do not maintain the same database platforms or produce identical student performance and administrative reports. Each system is customized to the particular district. This suggests an inherent problem if the ADE wishes to standardize student assessment and/or data tracking and reporting as part of a statewide accountability system.

Arizona districts are not at the same place and same time in terms of testing and data management capabilities. For example, Williams Unified District has collected last year's assessments to assemble in a district-wide reference book. The test items are not available in electronic version(s). Deer Valley District is developing opt out tests for all courses in the district, starting first with academics and then with electives. Scottsdale Unified District uses departmental tests. Chandler District uses teacher-prepared test materials and is currently maintaining a dual system (hand and electronic) for student competency attainment reporting. They anticipate crossing over to a full electronic system by next fall. Williams District is reviewing the VDMS computer tracking system for possible adoption. Mesa and Paradise Valley Districts maintain on-line information for accountability purposes. Clearly there are wide differences in district capabilities in terms of testing practices, data management, and performance accountability/reporting capability.

Whenever program/course or administrative requirements change, districts must expend large amounts of personnel, time, and money to affect the necessary change(s) across the entire instructional and management system. To the extent that reporting requirements and instructional content remain fairly constant, districts are able to maintain existing sophisticated systems and/or continue to move toward implementing systems.

To the extent that there are unexpected and/or increased state reporting requirements, the database support systems and capabilities are negatively affected. This suggests that ADE needs to maintain stable, consistent database requirements for accountability and performance assessment reporting, while districts transition to electronic formats and build staff and electronic capability.

Standards and Assessment

Nationally, employers, educators, labor organizations, and state agencies are concerned that students are not adequately prepared for entry-level jobs in their chosen career clusters. The National Association of State Directors of Career Technical Education Consortium recently reported that:

One of the key implications of the global marketplace is that employers throughout the United States want easily identifiable “coins-of-the-realm” assessments of the knowledge and skills an individual gains in any educational or training setting.⁴⁸ The five criteria used for coin-of-the-realm qualifications include being relevant, clear, achievable, measurable, and attractive.

Literature reviews suggest three types of standards utilized by states and districts are most commonly called academic, employability (workplace), and technical standards. The standards may be separate or consolidated into integrated statements that describe the specific use of academic or employability skills *in the context of technical tasks* (Bailey and Merritt 1995).⁴⁹

Standards represent what the teacher, district, and/or state expects a student to recall, replicate, manipulate, understand, and/or demonstrate prior to graduation. These standards represent expectations. Instructional assessments are designed to measure how close a student has come to meeting a standard—the expectation. In Arizona, to be adequately prepared, a student is expected to complete high school, earn a diploma, and possess academic, workplace and occupational knowledge and skills (technical skills).

To verify that each student has attained the knowledge and skills in academic, workplace and technical skills, school-based and state-sponsored testing and/or industry credentialing must be completed in Arizona. In the future, approved CTE program status will only be for programs that are employing assessments and/or certifications and that meet or exceed State Performance Measures discussed at a later point in this report. Inadequate performance results can lead to school sanctions and/or withdrawal of funding support for the career/technical education program(s).

Arizona Academic Standards

Academic standards cover traditional school subjects and are available from a variety of educational associations, consortia, and state education agencies. In Arizona, K – 12 schools test academic attainment of the Arizona Academic Standards using both a norm referenced test, the Stanford 9, and an Arizona Curriculum Standards aligned test, the Arizona Instrument to Measure Standards (AIMS).

⁴⁸Wills, Joan. *Promoting New Seals of Endorsements in Career Technical Education* The National Association of State Directors of Career Technical Education Consortium, Washington, DC. 2002. p.3.

⁴⁹ Bailey, T., and Merritt, D. *Making Sense of Industry-Based Skill Standards*. Berkeley, CA: National Center for Research in Vocational Education, 1995. (ED 389 897)

The Stanford 9 is a norm-referenced test used to compare Arizona student performance with students from all other states. The AIMS test is a state-developed testing instrument aligned with the Arizona Academic Standards that are the basis for instruction (K-12) in Arizona. Both tests are used as indicators of student academic achievement. Similar academic content standards and more formal accountability systems have now been adopted in all but one state.⁵⁰

Achieve, Inc, a private, not-for-profit organization, assists Governors and business leaders in development/implementation of high academic standards, assessments, and accountability systems. They maintain a National Clearinghouse database for researching academic standards. Other academic standards sources include the respective subject area (i.e. science, math) professional associations and their respective websites/resource materials, national professional associations such as ASCD and Phi Delta Kappa, and the National Educational Assessment Program (NAEP) offices in Washington, D. C. Most state education agencies, including Arizona, also maintain academic standards information banks for their respective states.

Technical Skills Standards: Vocational Competencies/Indicators

Curriculum frameworks in Arizona utilize competencies and indicators as organizers for the instructional content to guide instruction and to inform students of CTE program expectations. In this sense, they are comparable to the Arizona Academic Standards. The competencies include cross-program and program-specific (technical) competencies for both core and program career options.

A competency is defined as “an educational construct/concept derived from a workplace task, knowledge, skill or ability requirement.”⁵¹ Simply put, **competencies** tell learners what primary skills they will learn. Competencies reflect industry-approved knowledge, skills, and abilities needed by a work-ready employee. **Unlike skill standards, competencies do not reflect specific job duties and tasks.**

A well-written competency should:

- Match industry standards and meet with industry approval;
- Be appropriate for secondary level instructional programs;
- Reflect cognitive (knowledge), psychomotor (skills) and affective (attitudes) learning domains; and
- Represent higher order levels of development within those domains.

Indicators state performance and outcomes and tell the learner what he/she should be able to do as a result of a specific learning experience. They relate directly to the competency. If the learner can demonstrate performance leading to the outcomes specified by the indicators, the learner is said to have *mastered* the competency. **Indicators state outcomes, not specific instructional activities.**

⁵¹ Arizona Performance Measures. Secondary FY2002 Guidelines for Career and Technical Education Program Evaluation. (Revised January 2002).

Indicators should:

- State performance-based objectives;
- Describe specific outcomes that are measurable;
- Include higher-order knowledge, skills and attitudes; and
- Be an essential part of mastery of the stated competency.

A report completed for the Arizona Department of Education, CTE Division, in May 2001⁵² included eighteen (18) recommendations with respect to utilizing industry standards and assessments, incorporating higher-order skills in standards, and investigating district and industry-developed assessments for reference in the curriculum frameworks. Subsequently, many of the recommendations were initiated with newly implemented Curriculum Design Team makeup and procedures. This research project addresses information and strategies to implement several others.

Ten related recommendations from this prior report remain either partially or fully un-addressed and include:

Higher order skills:

1. Department and (curriculum) Design Teams need to assure that higher-order skills are integrated into the *existing* Frameworks. Future ADE documents should be developed based on an accepted list of experts. A supplemental list of these items should be given to CTE instructors to incorporate into their instruction.

Assessment:

2. A review team to look at existent assessment materials could be established to do a test-item analysis matching the items to Arizona competencies and to determine the appropriateness of other existent materials for Arizona use.
3. The review team should review other assessment items from local school districts to consider for state adoption.
4. Another option the Department could consider is developing a bank of “minimum” assessment items for each program to use. Use of the minimum test item bank could be a district option, with districts either accessing the item bank or certifying that their own assessments address the minimums or greater.

Accessing resources:

5. The Department should consider joining other consortia that specialize in vocational standards, curriculum, instructional materials, and assessment activities.

Business/industry materials

⁵² Norris, Carol and Croft, Vaughn. (2001) *Curriculum Design Process and Materials Format*, Arizona Department of Education, Phoenix, AZ.

6. At a minimum, there is a need to ... 4) match industry standards and select those appropriate to the instructional program, and 5) consider developing performance indicators/rubrics for assessment.

Data Management and Utilization:

7. Maintain a consistent appropriate data gathering and reporting process.
8. Design a process that more closely aligns district data collection efforts with state reporting requirements while some of the present performance data collection requirements are placed on hold temporarily. This recommendation applies both to enrollment and student assessment data collection efforts.
9. The Department, working with local districts, should examine and establish a methodology and reporting mechanism to track student completion of competencies from one program level to the next.
10. The Department and participating school districts will need to revise data collection methodologies and reporting practices to meet state monitoring requirements and, in turn, to adequately address federal reporting guidelines.

Workplace Skills, Cross-Program Competencies, and SCANS Skills:

Workplace Skills, sometimes called generic skills, are included in the Arizona Academic Standards, but they are not presently tested in the state assessment (AIMS). During the past two decades, the skills needed to succeed in the workplace have changed significantly. Technical skills remain important, but, increasingly, employers recognize that another category of skills are crucial for employees to work "smarter, not harder." These skills go by a number of labels including soft skills, core skills, non-technical skills, essential skills, generic skills and new basics.⁵³ Arizona and other states' Workplace Skills standards were developed because most students will spend more than a third of their lives in occupational endeavors and because of employer insistence, particularly in recent years.

Arizona Workplace Skills, included in the Academic Standards, are intended to integrate into the traditional curriculum, at all grade levels, with an emphasis on application of academics. The assumption is that workplace skills are developmental and encompass an individual's entire lifetime. In this context, "lifetime" is inclusive of all grade levels, K-14. The following Workplace Skills are included in the Arizona Academic Standards but not included in the present Arizona AIMS standardized test:

1. Students use principles of effective oral, written and listening communication skills to make decisions and solve workplace problems.
2. Students apply computation skills and data analysis techniques to make decisions and solve workplace problems.
3. Students apply critical and creative thinking skills to make decisions and solve workplace problems.

⁵³ Murnane, R. J., and Levy, F. *Teaching the New Basic Skills. Principles for Educating Children to Thrive in a Changing Economy*. New York: Free Press, 1996.

4. Students work individually and collaboratively within team settings to accomplish objectives.
5. Students demonstrate a set of marketable skills which enhance career options.
6. Students illustrate how social, organizational and technological systems function.
7. Students demonstrate technological literacy for productivity in the workplace.
8. Students apply principles of resource management and develop skills that promote personal and professional well-being.

Integrated curriculum, freestanding modules, and work-based projects address skills workers need in different jobs. The standards can be divided three ways to include:

- a. **Industry core standards** covering knowledge and skills needed in most occupations across a whole industry. These would apply to any worker in selected industries such as electronics, hospitality, or business. The Arizona Workplace Skills and Cross-Program Competencies typify this category.
- b. **Occupational family standards** covering knowledge and skills needed in a related set of occupations either in one industry or across industries. For example, this could include medical lab and radiography workers in the diagnostic cluster in the health care industry or data entry workers in any industry. Selected Arizona Workplace and Cross-Program competencies could apply in this category.
- c. **Occupationally specific standards** covering the specific knowledge and skills in a single occupation, as in a traditional CTE program technical skills requirements. The program competencies/indicators for each Arizona curriculum framework apply in this category. Workplace Skills and Cross-Program Competencies do not meet the criteria of “technical skills” and, therefore, do not apply in this category.

Richens and McClain surveyed 400 employers about their perceptions of workplace basic skills and competencies required for current and potential employees. The employers said that they wanted entry-level workers to possess employability skills *rather than technology competencies*. The most important attributes for employees to have (rating over 92.6%) were basic skills, thinking skills, personal quality skills, and interpersonal competencies. Technology competencies and systems competencies rated the lowest at 54.5% and 52.8% respectively.⁵⁴ In other words, for the most part most employers wanted SCANS and workplace skills representing industry core standards more so than technical skills for entry-level workers.

Changing workplace requirements, coupled with employer dissatisfaction with job applicants, led to efforts to define essential workplace skills for current and future employees. Recent researchers found the following skills mentioned most frequently: knowing how to learn; competence in reading, writing, and computation; effective listening and oral communication

⁵⁴ Richens, G. P., and McClain, C. R. "Workplace Basic Skills for the New Millennium." *Journal of Adult Education* 28, no. 1 (Summer 2000): 29-34

skills; adaptability through creative thinking and problem solving; personal management with strong self-esteem and initiative; interpersonal skills; the ability to work in teams or groups; leadership effectiveness; and basic technology skills.⁵⁵ [Six of the eight Arizona Workplace Skills mirror what these researchers cite in their frequency list.](#)

Generic job skills such as problem solving, reasoning, using judgment, and contributing ideas are considered essential workplace skills because “they are significant in a high-performance workplace.”⁵⁶ These are “look good” skills demonstrated through cognitive/affective ability, as compared to “work well” skills demonstrated through psychomotor ability (and referenced as technical skills).

SCANS Skills: In 1990, the Department of Labor’s Secretary’s Commission on Achieving Necessary Skills (SCANS) addressed workplace skills for entry-level employment. The SCANS panel identified three foundation skills and five competencies as desirable workplace skills.⁵⁷

Foundation skills embrace both academic and behavioral characteristics and were divided into three categories including:

1. Basic skills (reading, writing, speaking, listening, and knowing arithmetic and mathematical concepts);
2. Thinking skills (reasoning, making decisions, thinking creatively, solving problems, seeing things in the mind’s eye, and knowing how to learn);
3. Personal qualities (responsibility, self-esteem, sociability, self-management, integrity, and honesty).

Competencies defined by SCANS are divided into five domains including:

1. Resources (identifying, organizing, planning, and allocating time, money, materials, and workers);
2. Interpersonal skills (negotiating, exercising leadership, working with diversity, teaching others new skills, serving clients and customers, and participating as a team member);
3. Information skills (using computers to process information and acquiring and evaluating, organizing and maintaining, and interpreting and communicating information);
4. Systems skills (understand systems, monitoring and correcting system performance, and improving and designing systems);

⁵⁵ Clagett, C. A. *Workforce Skills Needed by Today's Employers. Market Analysis MA98-5*. Largo, MD: Prince George's Community College, Office of Institutional Research and Analysis, 1997. (ED 413 949) [op.cit.](#) and Oliver, K. M.; Russell, C.; Gilli, L. M.; Hughes, R. A.; Schuder, T.; Brown, J. L.; and Towers, W. "Skills for Workplace Success in Maryland: Beyond Workplace Readiness." In *Workforce Readiness: Competencies and Assessment*, edited by H. F. O'Neil, Jr. Mahwah, NJ: Lawrence Erlbaum, 1997.

⁵⁶ Bailey, T. and Merritt, D. *Making Sense of Industry-Based Skill Standards*. Berkeley, CA: National Center for Research in Vocational Education, 1995. (ED 389 897)

⁵⁷ Whetzel, Deborah, "The Secretary of Labor's Commission on Achieving Necessary Skills," Eric Digests ED339749. March, 1992. (http://www.ed.gov/databases/ERIC_Digests/ed339749.html)

5. Technology utilization skills (selecting technology, applying technology to a task, and maintaining and troubleshooting technology).

Seven of the eight Arizona Workplace Skills model the SCANS recommendations. The professional skills model (rather than technical skills model) curriculum framework requires integration of vocational and academic education because advanced generic skills such as those identified in the SCANS list are integrated with industry-related skills.⁵⁸ Research-based projects using industry standards, employability skills, integrated instruction, and performance assessments are being conducted at several sites. For example, Johns Hopkins University is presently involved in three closely related projects to examine curriculum integration: 1) SCANS 2000, 2) a general education assessment project, and 3) a program outcomes assessment project.

The SCANS 2000 project at Johns Hopkins University administers employability skills, diagnostic assessments, and a second assessment (task based) to students in the workplace or classroom. Test results are the basis for a student's individual development plan to address closing learning gaps. The university's efforts emphasize improving, documenting, and being accountable for student performance. Results are entered into an online Career Transcript so that if students move their transcript can follow them.⁵⁹

Arizona and other states are not unique in their effort to teach SCANS-type skills; many other states have adopted a variety of strategies for the teaching of generic skills. Canada, Australia, and the United Kingdom have also initiated similar programs to address generic skill development.⁶⁰

The list of skills being used varies across countries; however, most lists include communication skills, interpersonal and social skills, organization and planning skills, problem-solving skills, creative thinking, literacy, and technology skills. These are comparable to many of the SCANS skills/domains. The Australian key competencies add "cultural understanding" as a generic skill.⁶¹

Recent reforms and innovative programs, such as Tech Prep and High Schools That Work, incorporate "generic" skills as they offer students a rigorous academic background, technological literacy skill development, and learning experiences that are situated in the context of real-world environments.⁶² High Schools That Work emphasizes contextual learning and personal achievement.

⁵⁸ Lankard Brown, Bettina. *Skill Standards: Job Analysis Profiles Are Just the Beginning* Trends and Issues Alert ERIC/ACVE, 1997.

⁵⁹ *Ibid.* 5

⁶⁰ Lankard Brown, Bettina. *Generic Skills in Career and Technical Education Myths and Realities* No. 22, ERIC/ACVE. Washington, DC 2002

⁶¹ Werner, M. C. ***Australian Key Competencies in an International Perspective***. Leabrook, Australia: National Center for Vocational Education Research, 1995. (ED 407 587)

⁶² Pucel, D. J. "The Changing Roles of Vocational and Academic Education in Future High Schools." Paper presented at the Central Educational Science Research Institute, Beijing, China, October 4, 1999. (ED 434 242)

Work experience programs are an essential part of vocational/technical programs. Current learning theories support the teacher's role as one of facilitator (as in a work experience program) and not as lecturer or director. Contextual learning supports the notion that learning occurs as students develop knowledge, construct meanings, and test out their own theories in their community and social environments⁶³ that, in turn, support work place skills development and work experience programs.

Generic skills, used in combination with occupational or technical skills, are commonly "taught" as part of a work experience program giving students in job situations the opportunity to practice and consolidate the skills.⁶⁴ However, Guile⁶⁵ notes that because workplace experiences vary, learning opportunities are *not distributed equally* across them. Thus, "work experience has often ended up affirming the idea that its main purpose is to assist young people to learn how to reproduce pre-existing activities."

Although there is evidence that generic skills are being taught in schools, there is great ambiguity about what they are. Terms commonly used to describe them include key skills, core skills, transferable skills, personal transferable skills, and employability skills.⁶⁶

CTE Cross-Program Competencies:

In addition to the Academic and Workplace Skills, the Career and Technical Education Division (CTE) of the Arizona Department of Education has adopted nine cross-program competencies (core skills) to include in all Curriculum Frameworks. These nine include:

CROSS PROGRAM COMPETENCIES FOR VOCATIONAL PROGRAMS

1.0 DEVELOP AN INDIVIDUAL CAREER PLAN

- Investigate career options including entrepreneurship
- Develop career goals based on interests, aptitudes, and research
- Review/revise plan/goals on annual basis
- Manage personal and career goals
- Describe factors that contribute to job satisfaction and success

2.0 PREPARE FOR EMPLOYMENT

- Develop a résumé
- Complete job application process
- Demonstrate interviewing skills

⁶³ Lankard Brown, Bettina. *Generic Skills in Career and Technical Education Myths and Realities* No. 22, ERIC/ACVE. Washington, DC 2002

⁶⁴ Imel, Susan. (1999) *Work Force Education: Beyond Technical Skills* Trends and Issues Alert No. 1, ERIC/ACVE

⁶⁵ Guile, D. "Skill and Work Experience in the European Knowledge Economy." *Journal of Education and Work* 15, no. 3 (September 2002): 268-269.

⁶⁶ [Ibid.](#)

3.0 PARTICIPATE IN WORK-BASED LEARNING EXPERIENCES

- Use technology appropriate for the job
- Demonstrate positive work behaviors
- Demonstrate positive interpersonal behaviors
- Demonstrate safe and healthy work behaviors
- Adapt to changes in the workplace
- Participate in a variety of work-based experiences, i.e. paid or non-paid job

4.0 DEMONSTRATE ORAL COMMUNICATIONS SKILLS

- Conduct formal/informal research to collect appropriate topical information
- Use questioning techniques to obtain needed information from audience
- Interpret oral and nonverbal communications of audience
- Demonstrate active listening during communications
- Demonstrate appropriate technologies for a formal presentation
- Prepare and deliver presentations
- Deliver presentation incorporating both appropriate verbal and nonverbal communication techniques
- Communicate using equitable and culturally sensitive language for a diverse audience
- Demonstrate effective telephone technique

5.0 DEMONSTRATE WRITTEN COMMUNICATIONS SKILLS

- Conduct formal/informal research to collect appropriate topical information
- Organize information and develop an outline
- Write business communication using appropriate format for the situation
- Using appropriate technology, prepare draft document using established rules for grammar, spelling and sentence construction
- Utilize multiple technologies for written and presentation communications

6.0 EVALUATE THE ROLE OF SMALL BUSINESS IN THE ECONOMY

- Evaluate the role of small business on local, state national and international economies
- List the factors, including personal traits, which contribute to the success of small business
- Compare/contrast the advantages/disadvantages of sole proprietorships, partnerships and corporations
- Develop a business plan
- Conduct an employee needs analysis for the organization based upon a business plan
- Research business locations and equipment needs for the organization based upon the business plan
- Analyze the relationship of customer service and customer satisfaction on the success of a business.

**7.0 BUSINESS AND FINANCIAL MANAGEMENT PRACTICES
NEEDED FOR ENTREPRENEURS**

- Develop a budget based on an enterprise's business plan
- Develop an income statement for an enterprise
- Develop a balance sheet for an enterprise
- Interpret financial information for decision making and planning
- Monitor and adjust business operation based on financial performance
- Analyze insurance and benefit needs
- Analyze available banking services

- Describe the impact of quality business communications on the success of an organization
- Manage customer relations

8.0 EVALUATE LEADERSHIP STYLES APPROPRIATE FOR THE WORKPLACE

- Determine personal characteristics of effective leaders
- Compare/contrast leadership and management styles
- Describe how cultural/ethnic differences affect leadership styles within a group
- Describe how cultural/ethnic differences affect interpersonal interactions/communications within a group

9.0 PARTICIPATE IN LEADERSHIP ACTIVITIES SUCH AS THOSE SUPPORTED BY CAREER TECHNICAL STUDENT ORGANIZATIONS

- Determine the roles and responsibilities that leaders and members bring to an organization
- Evaluate characteristics of an effective team player
- Evaluate characteristics of effective teams
- Practice techniques to involve each member of the team
- Demonstrate team work
- Practice effective meeting management
- Participate in career development events
- Develop and implement a personal and professional improvement plan
- Demonstrate business etiquette
- Practice decision-making processes

CTE Cross-Program Competencies, SCANS, and Arizona Workplace Skills are similar, as shown in Table II on the next page. While they do not match item for item, the Arizona Standards clearly mirror and support the SCANS Foundation Skills and Competencies. In addition, the CTE Cross-Program Competencies mirror many of the SCANS and Arizona Work Place Skills. Table II provides a comparison of these similarities.

Table II: Comparison of SCANS, Arizona Workplace Skills, and CTE Cross-Program Competencies

SCANS Skills/Competencies	Workplace Skills	Cross-Program Competencies
Basic Skills (reading, writing, Speaking, listening, knowing arithmetic and mathematical concepts)	Oral, written and listening skills. Computation skills/ data analysis.	Oral communication skills. Written communication skills.
Thinking Skills (reasoning, solving problems, knowing how to learn)	Critical/creative thinking and decision-making.	
Personal Qualities (self-esteem, responsibility, self-management, sociability, integrity, and honesty)	Develop skills that promote personal and professional well being.	Prepare for employment.
Manage Resources (time, money, materials and workers)	Apply principles of resource management.	Participate in work-based learning activities. Demonstrate business/financial management practices for entrepreneurs.
Interpersonal Skills (diversity, leadership, team member)	Work within team settings.	Leadership styles appropriate for the workplace. Leadership/CTSO activities.
Information Skills (computers for information processing, interpreting and communicating info.	Illustrate how social, organizational and technological systems function.	
Systems Skills (understand, monitor,		

SCANS Skills/Competencies	Workplace Skills	Cross-Program Competencies
improve, design systems)		
Technology Utilization Skills (select and apply technology, maintain, troubleshoot technology)	Demonstrate marketable skills. Technological literacy for productivity in the workplace.	
		Develop an individual career plan.
		Evaluate role of small business in the economy.

The table cells above clearly show that two of the Arizona Cross Program Competencies (Develop an individual career plan and Evaluate the role of small business in the economy) do not cross-reference to either SCANS or the Arizona Work Place skills. In fact, “develop an individual career plan” is part of the process to “prepare for employment” and might be more appropriately combined with the “Prepare for employment” competency, rather than continued as a separate Cross Program Competency.

Many other states, including Virginia and New York, have adopted workplace and/or career standards. These two states’ standards are compared to Arizona Cross Program Competencies and are shown in Table III on the next page.

Table III:		
Arizona CTE Cross-Program Competencies	Virginia Workplace Readiness Skills	New York Career Development Occupational Studies (CDOS)
Demonstrate oral communication skills.	Demonstrate speaking and listening skills on a level required for employment in a chosen career field.	Students will demonstrate how academic knowledge and skills are applied in the workplace and other settings.
Demonstrate written communication skills.	Demonstrate writing skills on a level required for employment in a chosen career field.	Students will demonstrate mastery of the foundation skills/competencies essential for success in the workplace.
Participate in work-based learning experiences	<p>Demonstrate computer literacy on a level required for employment in a chosen career field.</p> <p>Demonstrate math skills on a level required for employment in a chosen career field.</p> <p>Demonstrate a strong work ethic.</p> <p>Demonstrate satisfactory attendance.</p>	
Evaluate leadership styles appropriate for the workplace.	Demonstrate understanding of the “big picture.”	
Develop an individual career plan.		Students will be knowledgeable about the world of work, explore career options, and relate personal skills, aptitudes, and abilities to future career decisions.
Participate in leadership activities such as those supported by CTE student organizations.	Participate as a team member to accomplish goals.	
Prepare for employment.	<p>Demonstrate reasoning, problem-solving, and decision-making skills.</p> <p>Demonstrate a positive attitude.</p> <p>Demonstrate self-presentation skills.</p> <p>Demonstrate independence and initiative.</p>	Students who choose a career major will acquire the career-specific technical knowledge/skills necessary to progress toward gainful employment, career advancement, and success in post-secondary programs.

Table III:		
Arizona CTE Cross-Program Competencies	Virginia Workplace Readiness Skills	New York Career Development Occupational Studies (CDOS)
Evaluate the role of small business in the economy.		
Demonstrate business and financial management practices needed for entrepreneurs.		

As shown in the table cells above, two CTE Cross-Program Competencies (Evaluate the role of small business in the economy and Demonstrate business and financial management practices needed for entrepreneurs) are not represented in either the Virginia or New York standards. It is, of course, probable that some other state(s) may include these two competencies in their standards. However, information from Table II and Table III suggest that “evaluate the role of small business in the economy” as a separate cross-program competency should be reconsidered.

Cross-Program Competencies and Workplace Skills Assessments

Cross-program competencies, along with Workplace Skills within the Arizona Standards, are required curriculum in each Arizona vocational program. If ADE chooses to have separate assessment requirements for these competencies, then sources listed in Table IV below could be considered. Many of these sources have also been referenced by the current Design Team reports located in Appendix F.

Table IV: Cross-Program/Work Place Skills Assessment Sources

Organization	Testing Instrument	Assessment Content
ACT WorkKeys® Job Profiling process.	WorkKeys ⁶⁷ Forty-one states are using WorkKeys to assess student employability and workplace skills by either purchasing an existing assessment or contracting with WorkKeys to develop a test that fits a state's or institution's needs.	<ul style="list-style-type: none"> • Applied Mathematics • Applied Technology • Listening Skills • Locating Information • Observation Skills • Reading for Information • Teamwork • Writing

⁶⁷ <http://www.act.org/workkeys/education/works.html> Web site for WorkKeys for Education describing eight foundational skills (skills needed to learn other skills).

Organization	Testing Instrument	Assessment Content
	Tests emphasize workplace application of skills, rather than academic applications.	
AlignMark	AccuVision Workforce Readiness System	<ul style="list-style-type: none"> • Customer Relations • Decision Making • Commitment to Quality • Personal Qualities • Responsibility • Self-esteem • Self-management • Sociability
Brainbench	Online assessment and certification of over 400 different skills that drive business success today.	<ul style="list-style-type: none"> • SCANS-type skills • Decision-making • Communication • Inter-personal skills
Educational Testing Service (ETS)	World's largest private educational testing and measurement organization. Library of 20,000 tests and measurement devices from the early 1900s to the present. Tests priced differently and include scoring and data analysis services.	<ul style="list-style-type: none"> • Personality traits • Inter-personal skills • Some SCANS-type • Attitude scales • General aptitude
Employability Skills Certificate Program Lifework Education Team Department of Public Instruction P. O. Box 7841 Madison, WI 53707-7841 Fax: 608-267-9275	Wisconsin Employability Skills Certificate Program Portfolio	<ul style="list-style-type: none"> • Basic SCANS Skills • Personal/Interpersonal Skills • Thinking/Information Processing Skills • Systems/Technology
Harcourt Brace	Differential Aptitude Tests, 5 th Edition and Career Interest Inventory. Workbooks "Guide to Careers" and "Guide to a Career Portfolio" complement the Differential Aptitude Tests.	<p>Tests/workbooks cover:</p> <ul style="list-style-type: none"> • Career Planning • Prepare for Employment • Written communication • Leadership activities
NOCTI (National Occupational Competency Testing Institute)	<p>Workplace Readiness (\$17.50 per student test)</p> <p>Note: An upscaled Workplace Readiness test is to be piloted in the spring, 2003 testing season.</p> <p>Dr. Mike Roberts, NAU, is the Arizona liaison and testing coordinator.</p>	<ul style="list-style-type: none"> • Individual Career Plan • Prepare for Employment • Work-based Learning Experience • Oral Communications • Written Communications • Small Business in the Economy (partial) • Leadership Styles (partial) • Entrepreneurship is <u>not</u> included.

Organization	Testing Instrument	Assessment Content
Ohio Occupational Competency Analysis Profiles (OCAPs)	Employability unit. The unit is included in all revised OCAPs and are available in paper-pencil or computer-delivered assessment.	<ul style="list-style-type: none"> • career development • decision making and problem solving • work ethic • job seeking skills job retention and career advancement skills • technology in the workplace • lifelong learning, economics education • balancing work and family • citizenship in the workplace • leadership • entrepreneurship
Oklahoma Department of Career and Technical Education http://www.odcte.ok.us MAVCC: Midwest Atlantic Vocational Curriculum Consortium http://www.mavcc.org (and) CIMC – Curriculum and Instructional Materials Center http://www.okcareertech.org/cimc	Department of Education is part of the Multi-state Academic and Vocational Curriculum Consortium (MAVCC) which develops and distributes competency-based instructional materials based on industry input. CIMC is a division of the Oklahoma Department of Career and Technology Education.	Competency-based. Uses performance tests with several performance levels. Instructional materials integrate SCANS skills and work place basics; OK does <i>not</i> have separate assessments for cross-program competencies.
Skills USA – VICA	Administered by NOCTI. Total Quality Curriculum emphasizes SCANS skills.	<ul style="list-style-type: none"> • Leadership Activities/CTSO Organizations • SCANS skills integrated into many competitive events.
V-TECS	Offers instructional use assessment items via a test item bank (unsecured site). Test banks include written and performance-based items. Workplace Skills CD Rom costs \$99.95 for members and non-members.	More than 600 test items for: <ul style="list-style-type: none"> • Developing an employability plan • Seeking and applying for employment opportunities • Accepting employment • Communicating on the job • Interpreting the economics of work • Maintaining professionalism • Adapting and coping with change • Solving problems and critical thinking • Maintaining a safe and healthy work environment • Demonstrating work ethics and behavior • Demonstrating technological literacy

Organization	Testing Instrument	Assessment Content
		<ul style="list-style-type: none"> • Maintaining interpersonal relationships • Demonstrating team work
WestEd Assessment Services in partnership with California Department of Education	Career Preparation Assessment. Guidelines for the career portfolio include skill areas and instructions for creating portfolio entries.	<ul style="list-style-type: none"> • Career development • Work samples • Performance-based skills assessment

According to the Education Commission of the States (2000)⁶⁸ Maryland is the only state that has established World of Work and Survival Skills for inclusion in the state’s core assessment system. While Arizona includes Work place Skills in its Academic Standards, current tests do not measure attainment in this area. By comparison, New Jersey requires students to pass an assessment and demonstrate skills in core areas identified by the State Board of Education. The New Jersey requirements focus on career/technical education programs, including workplace skills.

The 2002 National Association of State Directors of Career Technical Education Consortium reported that:

1. Georgia, Maryland, Tennessee and Virginia are developing tests comparable to those used in New Jersey.
2. Several other states (Florida, Georgia, Massachusetts, Ohio, Virginia, and West Virginia) use differentiated diplomas specifically addressing career technical education.
3. Many states use ACT Work Keys to assess work-readiness of students.
4. Several states have incorporated AP or IB achievements in their assessments and diploma-granting strategies.⁶⁹

Crosswalk: Arizona Academic Standards and CTE Cross-Program Competencies

An analysis of the Cross-Program Competencies, Workplace Skills and other Arizona Academic Standards was conducted. The crosswalk which includes math, science, and language arts is in Appendix D. It is possible that present Design Team recommendations may differ from crosswalk items in this report, largely because they did not use science standards. Thus, it is probable that ADE will need to conduct an analysis of the completed curriculum design team crosswalks to check for consistency with those contained in this report and modify, if necessary, the present framework recommendations.

Review of the crosswalk clearly demonstrates that the Academic Standards essentially already include the CTE Cross-Program Competencies. Assuming the crosswalk satisfies criteria of “adequacy and accuracy,” there is little, if any, need to develop additional

⁶⁸ ECS Clearinghouse Notes, Advanced Placement Courses and Examinations. (January 2000). Education Commission of the States, Denver, CO.

⁶⁹ Wills, Joan. *Promoting New Seals of Endorsements in Career Technical Education*. National Association of State Directors of Career Technical Education Consortium, Washington, DC 2002. 8

competencies/indicators within the respective vocational programs to address either Workplace Skills or Cross-Program Competencies. It seems more appropriate to address making the present ones work to assure that students acquire these skills prior to program completion.

National Standards and Assessments for Career and Technical Education Programs

Skill Standards: Technical skill competencies incorporate skill standards developed by trade and special interest associations, industry groups, state agencies and/or local entities. Voluntary skills standards been developed through the National Skills Standards Board (NSSB) authorized in 1994 by the Goals 2000: Educate America Act. The NSSB supports a *voluntary* national system of skill(s) standards. It utilizes a framework of career clusters within which skill standards can be developed and it supports partnerships of business, education, community and other stakeholders to develop the skill standards. A primary source of industry and vendor specific references can be found on pages 15-22 of the National Skill Standards Board website <http://www.nssb.org/certl.htm>

Maryland is the only state to include both academic and technical skills standards in its “core standards.” Other states have skills standards but differ in the degree to which technical and workplace skills are included in their core standards. Their standards are presented in uncommon formats and there is not consensus about what “soft skills,” academic, and/or vocational skills should be included in the standards.

Program administrators and teachers express concern about the extent to which local school personnel are able to adopt numerous academic and industry skills standards being developed at the national, state, and local levels.⁷⁰ They worry that restrictive budgets prohibit keeping current with industry technology uses. The availability of updated instructional equipment and materials to teach and incorporate technical practices and standards is also a concern.

As noted earlier, some states have adopted existing industry-based skill standards, or have embedded them in statewide assessment and certification programs (see page 14)⁷¹ Examples of this approach can be found in California’s Career-Technical Assessment Program (C-TAP); Ohio’s Integrated Technical and Academic Competencies (ITACS); and Oregon’s Certificate of Initial Mastery (CIM) and Certificate of Advanced Mastery (CAM).

Some states and local districts are using the National Skill Standards Board (NSSB) standards. However, NSSB projects (22) have been controversial because skill standards parameters have not been defined and the format and presentation process has not been standardized. Still, these materials are a primary reference for vocational educators and

⁷⁰ Ananda, S. M. et al. "Skills for Tomorrow's Workforce." *Policy Briefs*, no. 22. San Francisco: Far West Lab for Educational Research and Development, December 1995. (ED 392 132)

⁷¹ Rahn, M.L.; O'Driscoll, P.; and Hudecki, P. *Taking off! Sharing State-Level Accountability Strategies: Using Academic and Vocational Accountability Strategies to Improve Student Achievement*. Berkeley, CA: National Center for Research in Vocational Education, 1999. (ED 431 138)

business-industry representatives who encourage use of industry-based standards, assessments, and/or certification programs.

Industry Credentialing: Many states, including Arizona, are considering adopting/adapting industry-related credentials and/or vendor-specific certifications as competency attainment documentation. This may include company certificates, industry/trade certificates or diplomas, and state registrations, licenses, or certificates as indicators of student achievement. As a general practice, credentialing exams are given at the end of the respective program/course. This practice is not totally accepted among educators, largely because they believe assessment administered at the end of the learning process limits the contributions of other types of performance assessment.

Many educators feel that there should be a combination of intermittent assessments with an industry certification test used as a “capstone” experience. They believe that failure to pass the capstone test should not be viewed as failure to pass the program. Successful attainment of the program/course competencies, verified through a series of assessments, is considered the primary success indicator, rather than the capstone test in and of itself.

Other educators view industry-developed credentialing tests as the primary and, perhaps, only way to validate competency attainment. Reasons to adopt industry credentialing include aligning content, standards, competencies and assessment to nationally recognized industry accreditations. The Ohio Department of Education⁷² summarizes that doing so can lead to:

- Student credentials recognized by industry and employers nationwide.
- Standardized curricula so employers, colleges and apprenticeship programs can be confident students have received the same rigorous training regardless of where in Ohio they attended school.
- Industry/association/college partnerships with advanced credits for college or apprenticeship programs.
- Training based on industry-driven standards.
- Alignment to academics, assessment, apprenticeships and postsecondary admissions criteria.

From the viewpoint of industry, several obstacles threaten progress toward a certified national skill standards system. According to Geber,⁷³ skills standards will never succeed unless companies are convinced that they have something to gain. He notes that there is general consensus that national skills standards are or may be difficult to implement because:

1. Industry input is a time consuming process.
2. Occupations overlap job functions and industries.
3. Standards need to be fair and nondiscriminatory.
4. Geographic distributions and diverse interests make it difficult

⁷² <http://www.ode.state.oh.us/principal/assess/default.asp>

⁷³ Geber, B. "The Plan to Certify America." *Training* 32, no. 2 (February 1995): 39-42, 44.

to get people together to obtain agreement on standards.⁷⁴

Industry/vendor assessments and skill certifications: At the secondary school level, it is possible to use occupational skill certificates and state or industry-regulatory exams to acknowledge successful program completion. Regulatory exams result in registrations, licenses, and/or certifications. Regulatory credentials such as state registrations and licenses, industry/trade certificates or diplomas, credentials or certificates offered by industry associations, or unions, and vendor-specific company certificates (e.g., Cisco Certified Internet Expert) are representative examples of certifications.

There are some existing national standards, as in the hospitality industry and ASE automotive standards, but nationally not all instructional programs have adopted these respective standards. Some career/technical areas have adopted national skills standards such as those available from NSSB (described above) or industry associations (i.e. Pro Start certification in Food Production/Culinary Arts), but many have not. Many schools have adopted certain vendor-specific skill certifications for selected vocational programs such as A+ and CISCO in Business Administrative Services.

States such as Ohio, New York, Wisconsin, Indiana, Arkansas, and Virginia have been using vendor, consortium-developed, and/or industry-developed standards and credentials for some time. For example, Indiana and Arkansas use both V-TECS and industry-based certifications. New York makes extensive use of NOCTI tests for 12 of 17 trade areas and also uses industry tests/certifications.

The use of third-party assessments such as NOCTI for state licensure exams also has a long history in vocational education. Pennsylvania, Vermont, and Connecticut use NOCTI tests with secondary students for vocational skills assessment and certification. Arkansas uses NOCTI tests *if none are available* from an industry source. Additionally, more states are requiring students to apply for industry-sponsored credentials such as Automotive Service Excellence (ASE) and other industry-endorsed assessments.

Compared to many other states, industry/vendor credentials have had limited use in Arizona career/technical programs. These certifications are being reviewed in sequence with current Arizona state-sponsored curriculum development projects. Each new or revised curriculum framework presently incorporates industry-related credentials as one means of validating student achievement in the respective vocational program(s).

Copies of recently completed Design Team recommendations are included in Appendix E of this report. Also included is information from a previous University of Arizona assessment report (Fall 2001) summarizing potential certifications in other curriculum areas not yet addressed by the ADE curriculum teams. Curriculum Design Team resource lists in Appendix E were extensive and did not use consistent formats. Nonetheless, readers are

⁷⁴ Lankard Brown, Bettina. *Skill Standards: Job Analysis Profiles Are Just The Beginning* Trends and Issues Alert. ERIC/ACVE, 1997.

encouraged to review these Curriculum Design project certification/credentialing recommendations for use in their programs.

No attempt has been made to prioritize the certification(s) or rank order the assessment sources. This is a task will need to be accomplished as a next step in identifying “preferred” industry certifications to adapt/adopt for use in each program.

A formal adoption of the new ADE Curriculum Designs has not taken place in local districts, but, in the future, state-approved CTE program status in Arizona *will only be for programs employing approved assessments and/or certifications*. This means that each local district will need to identify and adopt the State model, select an alternative assessment model acceptable to the State, or select both State and local assessments under an ADE approved plan in order to maintain “approved program” status for accountability purposes in Arizona.

The Office of Vocational and Adult Education⁷⁵ reported on several surveys completed in 2001 related to the use of skill certificates and/or industry certifications. The surveys clearly show that there *is no standard practice* regarding the use of skill certificates and industry credentials for secondary and community college vocational/technical education students.

For example, the FRSS survey asked whether each secondary occupational program prepared students to earn either skill certificates or industry credentials. Seven percent of public secondary schools with listed occupational programs prepared students in *all* of their programs for a state or industry regulatory exam (leading to registration, licensing, or certification), while 41 percent prepared students in at least one of their programs to do so. Thirty-one percent of public secondary schools with listed occupational programs prepared students in *all* of their programs to earn an occupational skill certificate, whereas 55 percent prepared students in at least one of their programs to do so.

The second survey (PEQIS) asked community colleges about two standard academic credentials (associate’s degrees and institutional certificates/diplomas), about regulatory credentials (state registrations, licenses, or certificates), and two types of credentials offered by industry, associations, or unions (industry/trade certificates or diplomas and company certificates (e.g., Cisco Certified Internet Work Expert).

About half of less-than-4-year postsecondary institutions that offered at least one listed occupational program offered institutional certificates/diplomas in *all* of their programs. Eighty-seven (87%) percent offered this type of credential for at least one of their programs. Next most common were associate’s degrees and state-awarded regulatory credentials (registrations, licenses, or certificates), each offered by about half of these institutions for at least one of their programs. Industry/trade certificates or diplomas were available for at least one program at about one-third of these institutions, and company certificates were offered at about one-fifth of these institutions.

Many states have on-line comprehensive resource materials and lists of industry-endorsed assessment/credentialing materials they use. For example, the Ohio Department of Education website lists *comparable information for every vocational program/course* including

⁷⁵ US Department of Education, Office of Vocational and Adult Education, 2001. <http://www.ed.gov/ovae>

course/program information, licensure and license type/code, certification type/code, and course information. Course information includes career cluster, competencies, performance measures, and accountability indicators. A summary of the Ohio assessment program follows.

Ohio Department of Education Assessment Program

The Ohio Department of Education provides the Ohio Career-Technical Competency Assessments (OCTCA) in online or paper/pencil format for the following program areas:

1. Environmental & Agricultural Sciences
2. Business
3. Marketing Education
4. Family & Consumer Sciences
5. Secondary Health Careers
6. Secondary Industrial & Engineering Systems

Ohio utilizes skill standards testing and states that “All certifications/credentials are **both knowledge and performance based** and each is **nationally** affiliated.”⁷⁶ “Success Builds Ohio” is a statewide initiative launched in November 2000. It includes a construction and manufacturing focus and aligns content, standards, competencies and assessment to **nationally recognized industry accreditations**. A student certification/credential is conferred through Ohio accreditations in the form of a certificate or transcript. Examinations involving an Ohio certification/credential are ideally, but not inclusively, graded independent of the school. Four program and four student criteria for test selection are used and include:

Program accreditation is a method of determining through non-governmental peer evaluation that a school program meets or exceeds all established standards and requirements of academic/technical excellence in curriculum, student facilities, placement services, training facilities equipment, safety and instructor credentials. The purpose of accreditation to industry standards is to improve the quality of education and to establish a standard supported and developed by industry.

Self-evaluation process, a program accreditation requirement, takes into account objectives that include instructor qualifications, student-teacher ratio and safety issues. The evaluation also should provide an effective method for linking its trainees to potential employers and identifying advisory committee involvement.

An on-site validation process conducted through industry/ association joint participation verifies information on the self-evaluation process is correct.

Portable curriculum, included by some program providers, include standards, competencies and even lesson plans, while others work exclusively from task

⁷⁶ <http://www.ode.state.oh.us/ctae/principal/assess/default.asp>

lists, competencies or standards. In either case, each must be national in scope and portable in usability.

Four mandatory student criteria in the Ohio initiative include that the:

Examination(s) **confer(s) certification/credential.**

Examination(s) **graded independent of school.**

Certification/credential **knowledge and performance based.**

Examination(s) is/are **national.**

In addition to the career/technical competency assessments (OCTCA), Ohio maintains competency profiles (OCAPS) and a web site⁷⁷ listing all approved/recommended assessments with criteria and ratings, type/name of certification, test supplier, and contact information. The resource materials are organized by occupational area and, at present, are available for Family and Consumer Sciences and Industrial and Engineering Systems. A sample of the on-line resource document for Culinary Arts and Food Service Management follows.

Ohio: FAMILY and CONSUMER SCIENCES⁷⁸
CULINARY ARTS and FOOD SERVICE MANAGEMENT
Subject Code 09.0203

Program Description:

Preparation for employment in management, production and service positions in the hospitality and tourism industry

Provider must be national and meet a minimum of one criterion in each category below:

PROGRAM CRITERIA

1. accreditation by program
2. a self-evaluation process
3. on-site validation
4. portable curriculum

STUDENT CRITERIA

1. examination(s) confer(s) certification/credential
2. examination(s) graded independent of school
3. certification/credential knowledge and performance based
4. examination(s) is/are national

PROVIDERS	PROGRAM CRITERIA				STUDENT CRITERIA			
NRAEF, ACF	1	2	3		1		3	4
Pro-Start				4	1	2		4
*Pro-management				4	1	2		4

⁷⁷ [ibid](#)

⁷⁸ http://www.ode.state.oh.us/ctae/Ind_Std_Accreditation_Apprenticeships/default.asp

- Adult only

Provider:

National Restaurant Association Education Foundation (NRAEF)

Contact Information:

175 West Jackson Boulevard Phone: (312) 715-1010 or
Suite 1500 (800) 765-2122
Chicago, IL 60604-2702 Website: <http://www.nraef.org>

Student Certification:

- Pro-Start
- Serv Safe
- Pro-management – Foodservice Management Professional Certification (FMP)

Articulation:

- Food Service Industry

Provider:

American Culinary Federation (ACF)

Contact Information:

20 San Bartola Drive Phone: (800) 624-9458
St. Augustine, FL 32086 (904) 824-4468
Fax: (904) 825-4758
E-mail: acf@acfchefs.org
Website: <http://www.acfchefs.org>

Program Certification:

- Access ACF

Articulations:

- Food Service Industry

If Arizona wishes to implement industry assessment/certifications statewide, it should investigate establishing a database comparable to Ohio for teachers to access information about “acceptable” industry certifications in their program(s). Because of teacher turnover and the lag between updating the print versions of the Curriculum Frameworks, an online system would support industry-based credentialing efforts and be a valuable resource for districts. Implementing and maintaining an online resource of industry-approved certifications might necessitate a new support staff position within ADE.

Licensed Occupations: In Arizona, there are 63 licensed occupations regulated by a variety of licensing boards⁷⁹. Many of these licenses have age requirements (usually 18 and older) and some require additional education/training (i.e. Emergency Medical Technician). These requirements prohibit most secondary students from obtaining licensure, but should not be construed as prohibiting students from *planning to enter* the specialized areas. Thus, some secondary programs appropriately are pre-preparatory licensing, as in the case of certain cosmetology, emergency services, nursing, agri-science, and law, public safety and security occupations.

Table V: Arizona Licensed Occupations

License	Licensing Board
Accounting, Public - CPA	Accountancy, Board of
Advance Fee Loan Broker	Banking Department
Aesthetician	Cosmetology, Board of
Agricultural pest control advisor	Agriculture, Department of
Aircraft Dealers, Owners	Transportation, Department of
Applicators of pesticides	Agriculture, Department of
Aquaculture	Agriculture, Department of
Architect	Technical Registration, Board of
Automotive recycler	Transportation, Department of
Boiler	Industrial Commission of Arizona Arizona Division of Occupational Safety and Health
Boxer, trainer, promoter	Boxing Commission, Arizona State
Boxing Trainers, Judge, Referee, etc.	Boxing Commission, Arizona State
Chiropractors	Chiropractic Examiners, Board of
Citrus Fruit Broker, Dealer, etc.	Agriculture, Department of
Community College Teachers	Community Colleges, Board of Directors
Cosmetologist	Cosmetology, Board of
Dental assistant	Dental Examiners, Board of
Dental hygienist	Dental Examiners, Board of
Dentistry	Dental Examiners, Board of
Denture technology	Dental Examiners, Board of
EMT (Emergency Medical Technician)	Health Services, Department of Behavioral Health Licensure
Embalmer	Funeral Directors And Embalmers, Board of
Escrow agent	Banking Department
Funeral Directors	Funeral Directors And Embalmers, Board of
Handlers of dead or deceased live stock	Agriculture, Department of

⁷⁹ Arizona CareerInfoNet @ http://www.acinet.org/acinet/lois_agency.htm?stfips=04&by-state&x=28y=9

License	Licensing Board
Hay broker/dealer	Agriculture, Department of
Homeopathic Physicians	Homeopathic Medical Examiners, Board of
Investment advisor	Corporation Commission
Land surveyor	Technical Registration, Board of
Landscape architect	Technical Registration, Board of
Landscape architect, land surveyor	Technical Registration, Board of
Lobbyists	Secretary of State
Marriage and family therapist	Behavioral Health Examiners, Board of
Medical doctor, interns and residents	Medical Examiners, Board of
Midwife	Nursing, Arizona State Board of
Money Transmitter	Banking Department
Mortgage Banker	Banking Department
Mortgage Broker	Banking Department
Motor vehicle dealer	Banking Department
Nail technology instructor, salon, school	Cosmetology, Board of
Naturopathic medical practice	Naturopathic Physicians Board of Medical Examiners
Notaries Public	Secretary of State
Nursing	Nursing, Arizona State Board of
Optician services	Opticians Dispensing, Board of
Optometrist	Optometry, Board of
Osteopathic physicians and surgeons	Osteopathic Examiners in Medicine and Surgery, Board of
Packer or shipper, fruits and vegetables	Agriculture, Department of
Pesticide applicators users, sellers, and distributors	Agriculture, Department of
Pharmacists, pharmacy interns	Pharmacy Board
Physician assistants	Physician Assistants, Joint Board on the Regulation of
Podiatry	Podiatry Examiners, Board Of
Private Investigators	Public Safety, Department of
Private process servers	Supreme Court, Arizona Administrative Office of the Courts Court Services Division, Certification Unit
Psychologist	Psychologist Examiners, Board of
Radiologist	Medical Radiologic Technology Board of Examiners
Real Estate Appraiser	Appraisal, Board of
Real Estate broker and salesperson	Real Estate, Department of

License	Licensing Board
Security Guards	Public Safety, Department of
Substance abuse counselor	Behavioral Health Examiners, Board of
Taxidermy	Game And Fish Department
Teachers, Elementary & Secondary	Education, State Board of Certification Unit
Veterinary practice, technician	Veterinary Medical Examining Board
Vocational rehabilitation	Industrial Commission of Arizona Arizona Division of Occupational Safety and Health
Well drillers	Water Resources, Department of

Apprenticeships are another means of obtaining and certifying technical skills available in a variety of industries, but particularly in construction trade/technical areas such as carpentry and electrical. It is important to note that Arizona apprenticeship programs require a high school diploma or GED and the applicant must be 18 years of age prior to application for entrance into the apprenticeship. Apprenticeship programs are minimally 8,000-10,000 hours in length, full time, and with a 3-6 month probationary period depending on the apprenticeship area. Secondary students participating in a Tech Prep (2+2) program should be made aware of apprenticeship opportunities that may align with their curriculum specialization area⁸⁰ but also need to be made aware of unique education and pre-acceptance requirements.

Beyond the problem associated with age and pre-entrance criteria for secondary students wishing to enter apprenticeship training, some Arizona educators are concerned that the extensive on-the-job training may not align with the secondary schools' program competencies. An additional concern is that related apprenticeship instruction may be excellent, but equate to a relatively short period of time.


Experience has shown that the unions are reluctant to recognize related instruction provided in secondary schools and often will not give credit for prior learning. Provided instruction is equal to or exceeds that offered through the apprenticeship program, CTE completers entering into apprenticeship programs should take less time to attain journey status. However, this is not generally the case. This does not mean that secondary students should be counseled away from apprenticeships; instead they should be made aware of potential "loss" of related secondary program experience(s).

Thirty-five apprenticeship programs are in Arizona and summarized in Table VI that follows.

⁸⁰ Available through the Arizona State Apprenticeship Council.

Table VI: Arizona Apprenticeship Programs

Industry	Occupation	Location	
Arizona Asbestos Workers JATC	Construction	Insulation Worker	Maricopa County
Arizona Builders Alliance	Construction	Concrete Form Builder	Maricopa County
Arizona Builders Alliance	Construction	Electrician	Maricopa County
Arizona Builders Alliance	Construction	Pipefitter	Maricopa County
Arizona Builders Alliance	Construction	Plumber	Maricopa County
Arizona Builders Alliance	Construction	Sheet Metal Worker	Maricopa County
Arizona Builders Alliance	Construction	Sign Erector	Maricopa County
Arizona Chapter, Associated General Contractors	Construction	Equipment Operator	Maricopa County
Arizona Child Care Apprenticeship Committee	Service	Childcare Development Specialist	Maricopa County
Arizona Concrete Contractors Association	Construction	Cement Mason	Maricopa County
Arizona Concrete Contractors Association	Construction	Concrete Form Builder	Maricopa County
Arizona Laborers Apprenticeship	Construction	Craft Laborer	Maricopa County
Arizona Masonry Contractors Association	Construction	Bricklayer	Maricopa County
Arizona Operating Engineers JA&TS	Construction	Mechanic, Construction Equipment	Pinal County
Arizona Operating Engineers JA&TS	Construction	Operating Engineer	Pinal County
Arizona Operating Engineers JA&TS	Construction	Plant Operator	Pinal County
Arizona Precision Sheet Metal, JIT	Construction	Press Brake Operator	Maricopa County
Arizona Public Service Company JAC	Utilities & Transportation	Electrician	Maricopa County
Arizona Public Service Company JAC	Utilities & Transportation	Lineman	Maricopa County
Arizona Public Service Company JAC	Utilities & Transportation	Meter Repairer (Any Kind)	Maricopa County
Arizona Roofing Industry JATC	Construction	Roofer	Maricopa County
Arizona State Carpenters (SE Area)	Construction	Carpenter	Pima County
Arizona State Carpenters JA&TC	Construction	Carpenter	Maricopa County
Arizona State Carpenters JA&TC	Construction	Lather	Maricopa County
Arizona State Carpenters	Construction	Millwright	Maricopa County

Industry	Occupation	Location	
JA&TC			
Arizona State Carpenters JA&TC (Northern Area)	Construction	Carpenter	Coconino County
Arizona State Carpenters JA&TC (Northern Area)	Construction	Lather	Coconino County
Arizona State Carpenters JA&TC (Northern Area)	Construction	Millwright	Coconino County
Arizona State University	Service	Glassblower	Maricopa County
Arizona Teamsters JA&TS	Construction	Truck Driver, Heavy	Maricopa County
Arizona Tire & Service Dealers Association	Service	Mechanic, Automobile	Maricopa County
Arizona Tire & Service Dealers Association	Service	Undercar Specialist	Maricopa County

Academic and Technical Competency, Assessment, and Certification

In all programs, students are expected to attain the Cross-Program Competencies and Workplace Skills on completion of a Level III program. Academic competencies that serve across the career cluster areas are integrated with technical skills within the occupational cluster. By design, there may be a particular focus that leads to industry credentialing or state/association licensing, tech prep articulation agreements for college credits, and/or continuation in a college program to attain an Associate Degree.

Drummond, Nixon, and Wiltshire⁸¹ offer three broad approaches on how to develop various types of skills within the curriculum. They include:

- Integrate generic skills within the career-technical education curriculum.
- Use free-standing modules that are not integrated into the curriculum, relying on the support of student tutors.
- Initiate work placements or work-based projects that will help students to develop employment-related skills within the context of real-world situations.

Depending on the particular CTE program, several program options may be available to students. For example, Arizona is presently reviewing the Law Enforcement curriculum and will validate program options. Among materials under review is the following list of program options as used in Kansas for their Law Enforcement cluster:

- Administration of Justice
- Criminal Justice/Police Science/Corrections
- Diver (Professional)
- Emergency Dispatcher
- Fire Control Technology
- Fire Science Protection Technology

⁸¹ Drummond, I.; Nixon, I. and Wiltshire, J. (1998) *Personal Transferable Skills in Higher Education: The Problems of Implementing Good Practice*. Quality Assurance in Education No. 1 19-27, 21

- Legal Assisting/Paralegal
- Legal Secretary
- Police Academy⁸²

There are nine options in the Kansas cluster. Theoretically, each program option represents a potential area for integrated curriculum frameworks and instructional practices leading to industry/association/vendor certifications and/or state licensing. The wider the range of program options, the longer the potential certification list becomes. In the case of Arizona, Cross-Program Competencies and Workplace Skills are common to all vocational/technical clusters and could utilize a common assessment. In addition, some clusters may have more/less academic skills appropriate to the cluster and each cluster will have varying numbers of technical skills. For this reason, each vocational program selects assessment strategies appropriate to the additional academic and technical competencies to be validated within each program option.

There is the potential for each program and each program option to have several minimum required assessments for Academic Standards. For example, a Legal Assistant/Paralegal would have academic performance standards somewhat different from those required for an Emergency Dispatcher, because the paralegal would require higher writing skills than the dispatcher. Conversely, the dispatcher would require higher oral presentation/speaking skills than the paralegal might need. In workplace skills, the dispatcher might require less interpersonal skills than the paralegal, but both might require equal information-handling and organizational skills.

The issue is *how to use a common assessment for these differing expectations*. The differences in programs and what assessment to use will require extensive review and consideration of existing assessments, development of new assessments and/or compromise between CTE programs to select a common assessment that most meets each of their respective needs to assess competencies and acceptable performance.

Current curriculum design teams in Arizona have been/are producing program competencies and recommended assessments including potential CTSO events and industry/vendor certifications. Materials in this report summarize Design Team industry/vendor certifications and those from other sources (Appendix E) that may be useful to future teams. Technical skills competency assessment sources are presented in Table VII on the next page.

⁸² It is important to note that Law Enforcement jobs are limited to age 21 and older. A student may enter a Law Enforcement training academy at age 19 with the understanding that they will qualify for employment upon graduation and reaching age 21.

Table VII: Academic/Technical Skills Assessment and Certification Sources		
Organization	Testing Instrument and Type	Assessment Content
AlignMark http://www.alignmark.com	The <i>AccuVision Systems</i> evaluate a candidate's skills and abilities that are required for success in a specific job position.	The <i>AccuVision Workforce Readiness System</i> is a unique assessment tool that uses job simulation, video and computer technologies to capture the skills and abilities required for success in customer care and a variety of customer contact, entry level positions. Skills assessed include: Customer Relations, Decision Making, Commitment to Quality, Personal Qualities, Responsibility, Self-esteem, Self-management, and Sociability.
Arizona Community Colleges	Tech Prep Articulation; college course syllabi; college course assessments per instructor.	Occupational specific content derived from the approved program and college articulation agreement that identifies specific course content.
Arizona Department of Education	Recommended in ADE curriculum design projects and subject to ratification.	Competencies/indicators and assessments/certifications per curriculum framework. May include some cross-program and/or workplace skills also.
Arizona State Government, Licensing and Credentials (see agency list below)	Testing instrument varies with licensure/credentialing agency. Guidelines available from each agency (i.e. Dept. of Health)	Age requirement and background check required for many licenses. Technical skills per agency guidelines.
California Department of Education: Career Technical Assessment Project (C-TAP) developed by Far West Laboratory for the State of California. http://www.cde.state.ca.us	Uses cumulative and administered assessments. Cumulative assessments include supervised practical experience, an assessment project, and a portfolio of work. Administered assessments have structured exercises, including project presentations, written scenarios focusing on solving a technical problem within the vocational area, and an on-demand test.	Arranged by instructional content areas, including vocational/technical.
Departments of Education for Florida, Georgia, Massachusetts, Ohio, Virginia, and West Virginia	Each state currently awards <i>differentiated diplomas</i> specifically addressing career technical education	See specific state web site for diploma requirements.
Departments of Education for Kentucky, Maryland, Michigan, Colorado, and Missouri	Each state currently has <i>academic and vocational/technical standards</i> on line.	See specific state web site for standards, assessments, etc.
Indiana Department of Education:	Proficiency guides available for seven	Manufacturing, academic &

Table VII: Academic/Technical Skills Assessment and Certification Sources		
Organization	Testing Instrument and Type	Assessment Content
Indiana Essential Skills and Technical Proficiencies Initiative http://www.state.in.us/dwd	vocational program areas. Indiana uses V-TECS and industry-based certifications. Brochure on industry-based certificates for workplace certification is available on-line. Scenario assessment software is available.	technical skill standards. Proficiency guides for: bio-science, business support, electronics, health, metal-working, plastics, and printing.
LJ Technical Systems http://www.ljtechnicalsystems.org	Custom development available, fee based. Uses computerized assessment and tracking system.	Instructional modules for 25 areas, including employability skills, customer relations and entrepreneurship.
Local School District (i.e. Glendale HS District, Tempe Union)	Teacher/district developed, continuous assessment program. Database for performance results and program reporting.	Criterion-referenced testing; may be pencil and paper or performance tests with scoring rubrics/guidelines.
National Healthcare Skill Standards http://www.mhc.org and http://www.nchste.org (HCSS)	National consortium working with NOCTI to develop online assessment for the Health Science Cluster. Available winter-spring 2003.	Knowledge and skills standards include: academic foundations, communication, systems, employment skills, legal responsibilities, ethics, safety practices, teamwork, health maintenance practices and information technology.
National Skill Standards Board (NSSB) http://www.nssb.org	Based on technical skill standards for entry-level positions. The NSSB has categorized the workforce into 15 industry sectors. NSSB uses common language format and guidelines for writing employability, academic, and occupational/technical <i>standards</i> (not skills). Format and presentation process have not been standardized. Employability skills areas closely mirror Cross-program Competencies and Arizona Workplace Skills. Provides on-line resource list of Certification/Apprenticeship standards for most vocational programs.	Employability skills include: listening, speaking, analyzing and solving problems, making decisions and judgments; organizing and planning; using social skills; using information and communications technology, gathering and analyzing information, working in teams; leading others, building consensus and self/career development. Employability components include critical functions of the job, key activities, and performance indicators.
New York State Education Department Office of Workforce Preparation and Continuing Education http://www.emsc.nysed.gov/workforce/cte/nationals (or) http://www.owpce.state.ny.us	Career Development and Occupational Studies (CDOS) includes learning standards. State uses industry-developed tests/credentials and NOCTI tests for 11 of 16 vocational program assessments. Two areas use state licensing agencies.	Sixteen “trade area” programs are listed with the name of the test/license required and the sponsoring national and/or state organization.
NOCTI (National Occupational Competency Testing Institute) provides Job Ready Student	NOCTI is a leading provider of occupational competency assessments and services. NOCTI's	71 standardized technical tests in occupational fields or customized assessments for

Table VII: Academic/Technical Skills Assessment and Certification Sources		
Organization	Testing Instrument and Type	Assessment Content
Occupational Competency Achievement Testing (SOCAT) http://www.nocti.org	products and services include job and task analysis, test development, written and performance assessments, scoring services and specialized reporting. Assessments in 14 occupational areas. Tests include written and performance parts. Over 70 standardized technical tests are available. Skills certificates are awarded to recognize attainment. Tests can be customized using test items from nearly 700 duty areas NOCTI has developed. Dr. Mike Roberts, NAU, is the Arizona liaison and testing coordinator.	specific site needs. Skills-USA test is administered by NOCTI. Workplace Readiness assessments for: communications, work-based learning, career plans, preparing for employment. Job Ready assessments for 34 of 36 Arizona vocational/ technical programs including: Agriculture Business Related Computer Related Construction Trades Consumer Economics Culinary Arts Drafting Electrical/Electronics Health Related Heating/Air Conditioning Machine Trades Maintenance Services Media Services Transportation
North Central Association http://www.nca.org	"Transitions" program is a self/peer-review process to develop student improvement plans for individual sites.	Assesses basic skills, employability skills, reasoning and information processing, problem solving and critical thinking, and career planning.
Ohio Department of Education, Career and Technical Education Division http://www.ode.state.oh.us/ctae/principals/assess/default.asp	Ohio Career Technical Competency Assessment (OCTCA) Test Crosswalk (See sample on page 49)	Program is competency based. Testing is both knowledge and performance based and uses both on-line and paper/ pencil formats. Covers six program areas. Lists nationally affiliated tests and cooperating groups.
Ohio Department of Education, Career and Technical Education Division http://www.ode.state.oh.us/ctae/principals/assess/default.asp	Integrated Technical and Academic Competencies (ITAC). ITAC scenarios are problem solving or performing tasks to demonstrate knowledge and skills in context. Project-based and other learning activities are provided to compliment ITAC scenarios for teacher use. Emphasizes high level academics and integrated curriculum.	Math, Science, Language Arts, Social Studies, Arts, and Foreign Language connections, as appropriate, are identified for each scenario. Some link directly to Ohio's Academic models.
Ohio Department of Education: Occupationally Competency Analysis Profiles (OCAPS) (http://www.ode.state.oh.us/ctae)	Competency analysis profiles for career and technical education. Each OCAP identifies occupational, academic, and employability skills. Uses career cluster competencies and sample work-place scenarios for	<u>Employability unit</u> includes: career development, decision making and problem solving, work ethic, job seeking skills, job retention and career advancement skills,

Table VII: Academic/Technical Skills Assessment and Certification Sources		
Organization	Testing Instrument and Type	Assessment Content
Ohio Department of Education: Occupationally Competency Analysis Profiles (OCAPS) continued:	assessment. Employability unit topics mirror AZ Cross-Program Competencies and Workplace Skills. Each competency list has two levels of items: core and advancing. Core items are the basis for questions on the Ohio Vocational Competency Assessment (OVCA).	technology in the workplace, lifelong learning, economics education, balancing work and family, citizenship in the workplace, leadership, and entrepreneurship. OCAP profiles exist for: agricultural, health, trade/industrial education, agribusiness, business/marketing, family/consumer sciences, dropout prevention, <u>and</u> applied academic programs (communications and mathematics).
Oklahoma Department of Career and Technical Education http://www.odcte.ok.us MAVCC: Midwest Atlantic Vocational Curriculum Consortium http://www.mavcc.org (and) CIMC – Curriculum and Instructional Materials Center http://www.okcareertech.org/cimc	Department of Education is part of the Multi-state Academic and Vocational Curriculum Consortium (MAVCC) which develops and distributes competency-based instructional materials based on industry input. The Curriculum and Instructional Materials Center (CIMC) is one of the nation's largest developers of competency-based instructional systems. CIMC is a division of the Oklahoma Department of Career and Technology Education.	Competency-based. Uses performance tests with several performance levels. Training and competency profiles crosswalk to national standards. Objectives, instructional strategies, and assessments are provided for technical skills areas in all occupations. Instructional materials integrate SCANS skills and work place basics; OK does <i>not</i> have separate assessments for cross-program competencies.
Putnam Valley Schools Index of Standards http://www.putnamvalleyschools.org	Resource site with links to all other states and many consortiums. Standards are sorted by subject area.	Standards for occupational, business, and technology.
Riverside Publishing http://www.uiowa.edu or http://www.riverpub.com	Achievement tests (ITBS) and intelligence tests only. Limited applicability for cross-program competencies.	ITBS includes information sources (reference publications and organizing information) and some math.
The Mackey Group www.mackeygroup.com	Fee-supported business/industry group developing performance-based skills standards for nine areas. Competency skills assessment is certified through ACT's Work Keys system (see Work Keys description below). Materials largely developed for post-secondary programs. Standards being developed for health/human services and marketing/PR should be reviewed	Includes foundational and personal workplace skills. Skills standards for allied dental health, cosmetology, secondary wood products manufacturing, public safety, & chiropractic. In process for: health and human services, food processors, plastic and

Table VII: Academic/Technical Skills Assessment and Certification Sources		
Organization	Testing Instrument and Type	Assessment Content
	for appropriateness to secondary vocational programs.	reconstructive surgery, and marketing/PR.
Vocational Evaluation and Work Adjustment Association (VEWAA) http://www.vewaa.org	Specializes in occupational/technical vocational assessment; website links to provider services.	Vocational evaluation and assessment/appraisal process to identify an individual's vocational potential. Includes work attitudes, skills, etc.
V-TECS(Vocational-technical Education Consortium of the States)	Offers instructional assessment items via a test item bank. The test banks include both written and performance-based items. These items are not generally used for pre or end-of-program competency assessment. V-TEC materials can be modified to fit local needs and, thus, do not represent secure tools for large-scale assessments. Used extensively in some states or as an option in others (i.e. Arkansas)	Arranged by instructional content areas. Includes some workplace and "soft skills" assessment items including: communications, maintaining professionalism, interpersonal relationships. Demonstrating teamwork, work ethics and behaviors. Developing an employment plan. (see Table IV for additional detail.)
Wisconsin Department of Education http://www.wde.wi.us	Employability Skills Certificate Program	Assessments for basic SCANS skills, thinking/information processing skills, and personal/interpersonal skills.
Wisconsin Technical College System Foundation Worldwide Instructional Design System (WIDS) http://www.wids.org	Provides software, professional development and training including assessments (rubrics and checklists, learning modules, course syllabi, program profiles and DACUM occupational profiles. Uses a computer-based tracking system for student performance reporting.	Purchased services including existing instructional materials and assessments or WIDS will custom design materials for school sites.

There are existing assessments and resource materials from other state education agencies, consortiums, universities and/or private vendors that appear to have applicability to technical skills, cross-program and workplace skills assessment. Design Teams should continue to access these resources and districts could explore purchase of available existing tests. No single test is particularly expensive in and of itself, but if districts were to purchase tests for each student, the costs could be prohibitive given the current state of school budgets. Conversely, for per student costs to be absorbed by ADE there would need to be a significant re-allocation of resources to support such purchases.

ASSESSMENT TYPES AND INSTRUCTIONAL USES

Stecher⁸³ provides four categories of alternative assessment to include:

⁸³ Stecher, et al (1997) "The Cost of Performance Assessment in Science: The Rand Perspective" Paper presented at the Annual Meeting of the National Council on Measurement in Education San Francisco, CA (April 1995) ERIC No. ED 383 732.

1. Written tests which may be multiple-choice, essay and/or writing samples.
2. Performance tasks
3. Senior projects such as research papers, performance projects and oral presentations. These projects have criteria and performance indicators, set standards and criteria in advance, and scoring rubrics for each project.
4. Portfolios

Ensuing portions of this report contain descriptions of several assessment strategies, pro/con statements regarding the strategy, and criteria for selection of the strategy. Appropriate references are contained in Appendix G: Authentic Assessment and Test Writing Skills Articles and are noted in the narrative. For example, several resources to help teachers develop test items are noted and include both technical requirements as well as analysis formulas to validate test items.

Objective and Subjective Test Item Construction

Developing test items is a process⁸⁴ that includes:

1. Specify the domain of knowledge and understanding that the student is required to learn.
2. Identify the mental tasks or processes that the student must use in dealing with the particular subject matter (recall, analysis, generalization, application, and discovery).
3. Write test questions or items that unambiguously assess the student's ability to deal with the knowledge domain as described by Steps 1 and 2.

Writing good test items is a time-consuming and difficult task. Gronlund⁸⁵ and Haladyna⁸⁶ are excellent sources to use regarding test design and writing test questions. All test items may be described as either objective or subjective. Objective items require students to select the correct response from several alternatives and may include multiple-choice, true-false, matching and completion item forms. Objective tests have a "content" that is subjective, because the content areas to be tested are selected by the instructors and/or test developer. Scoring methodology of the test is "objective;" thus, the test type is described as objective even though content selection is subjective.

Subjective items (often called essay items) permit students to organize, develop, and present information in an original written answer form and include short-answer essay, extended-response essay, problem solving and performance test item types. These forms are summarized in the materials that follow and include criteria for selection.

MULTIPLE CHOICE TESTS/TEST ITEMS

⁸⁴ Penn State University Testing Services (2000) Academic Testing Test Design and Construction. http://www.uts.psu.edu/Test_construction_frame.htm

⁸⁵ Gronlund, N. E. (1982) Constructing achievement tests. 3rd ed. Englewood Cliffs, JM: Prentice-Hall.

⁸⁶ Haladyna. T. M. (1999) Developing and validating multiple-choice test items. 2nd ed. Mahwah, NJ: Erlbaum.

Most objective tests are constructed using a multiple-choice test item format because they are perceived as fairly easy to construct. They "test" (measure the acquisition of) "facts" (knowledge) more than "higher order skills" (i.e. thinking and/or reasoning). Both memory and reasoning skills can be tested through appropriately written test items, but items that measure reasoning skills are difficult to construct because of the complexity of that particular domain (i.e. reasoning or critical thinking).

PRO

1. Multiple-choice tests/test items have a wide body of data showing that they "can be useful predictors" of job performance and college grades, if they are properly developed and interpreted (National Academy of Sciences Committee Conclusion, 1982.)
2. Objective tests usually measure knowledge of facts more efficiently than other test types, such as essay and portfolios.
3. Items can sample a wide range of content or objectives.
4. Multiple-choice test items can be utilized in test item banks effectively.
5. Test items are versatile, efficient, and have scoring accuracy, efficiency, and economy.
6. Provide an objective measurement of student achievement or ability and highly reliable test scores.
7. Patterns of incorrect responses can provide individual and group learning diagnostic information.
8. Guessing is reduced when multiple-choice items are used rather than true-false items.
9. Well constructed multiple-choice items can also be used to construct more realistic true-false items.
10. Item analysis is easily conducted on multiple-choice items which can then be improved before re-use.

CON

1. Test items are difficult and time consuming to write (particularly plausible distractors). Test quality is dependent on the item-writing skill of the instructor.
2. Factual knowledge rather than higher-level skills and understandings are frequently the basis of test items.
3. Better students are penalized and not given the opportunity to demonstrate extended knowledge.

4. The ability of multiple-choice tests to predict a student's future performance inside or outside of the classroom is far from perfect.
5. Multiple choice tests do not measure many traits that are important for success in the workplace or in further education, such as persistence and the ability to organize time and work.
6. Multiple-choice test items are often biased against minorities, economically disadvantaged students and women. They may not be culturally or equitably free/fair.
7. Standardized multiple-choice tests have drawn increasing fire for being too simplistic and not adequately measuring a student's ability to think and solve problems. ([New York Times](#))
8. Place a high degree of dependence on the student's reading ability and the instructor's writing ability.

Criteria for Multiple-choice Items

Three resource sites providing excellent guideline sources for multiple-choice test item writing can be found at <http://ericae.net/pare/getvn.asp?> (web site for the ERIC Clearinghouse on Assessment and Evaluation), <http://www.oir.uiuc.edu/dme/exame/ITQ.html> (web site for the University of Illinois Urbana-Champaign and at <http://www.use.umn.edu/oms/multchoice.htmlx> (web site for the University of Minnesota, Office of Measurement Services).

A reprint of the Penn State University Testing Services article provides analysis guidelines for test item construction, reliability, and validity. Also included is an Item Writing Guidelines article prepared by John P. Sevenair at Xavier University⁸⁷ and a resource guide from the University of Illinois Urbana-Champaign Office of Instructional Resources.⁸⁸ The articles are titled *Writing Multiple-Choice Test Items*, *More Multiple-choice Item Writing Do's and Don'ts*, and *University of Minnesota Writing Multiple-Choice Items* and are located in Appendix G.

In general, the authors⁸⁹ suggest that test items should be:

- Written clearly and succinctly

⁸⁷ Sevenair, John P., Item Writing guidelines, Xavier University, <http://webusers.xula.edu/jsevenai/objective/guidelines.html>

⁸⁸ University of Illinois Urbana-Champaign "Improving Your Test Questions" Urbana, IL. <http://www.oir.uiuc.edu/dme/exams/ITQ.html>

⁸⁹ Kehoe, Jerard (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation*. 4(9) <http://ericae.net/pare/getvn.asp?v=4&n=9>; or ERIC ED398236 http://www.ed.gov/databases/ERIC_Digests/ed398236.html; Frary, Robert B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation*. 4(11); <http://ericae.net/pare/getvn.asp?v=4&n=11>. or ERIC ED 398238 ERIC Clearinghouse on Assessment and Evaluation, Washington, D.C.; and University of Minnesota, Office of Measurement Services (1999) "Writing Multiple-Choice Items" <http://www.ucs.umn.edu/oms/multchoice.htmlx>

- Present a definite, explicit and singular question or problem in the stem.
- Represent answers that are unequivocal.
- Use distractors that are plausible competitors for the right answer (all alternatives are plausible and attractive to the less knowledgeable or skillful student).
- Be developed over time and based on “expert” editorial review
- Represent specific content areas.
- Deal with *significant* facts or concepts; measure relevant skills/knowledge.
- Not contain cues or clues leading to the right answer.
- Utilize appropriate stem, option, and distractor writing guidelines.
- Provide options that are mutually exclusive and not overlapping.
- Use parallel writing style and form.
- Avoid absolute language (“never” and “always”) and use “none” and “all of the above” sparingly.
- Use 3-5 options listed in logical order (if there is one); this approach lowers the probability of getting the item correct by guessing.
- Distribute the correct option randomly among the option positions.
- Consider as distractors correct responses that do not answer the question posed by the stem.
- Include as much information in the stem and as little in the options as possible.

TRUE/FALSE, MATCHING, AND COMPLETION

True/False items are sometimes useful in spotting misconceptions or knowledge gaps in subject fields. T/F items are susceptible to guessing by students due to the limited alternatives provided within the question format. One way to overcome guessing is to construct items that a variety of choices, usually two choices plus a conditional completion response. This type of item construction is called a compound form.

Matching test items are best used for recognition/association or labeling of terminology and concepts. Reasonable distractors are used to improve selectivity in responses.

Completion items are useful in measuring recall and general subject comprehension; they encourage less guessing than T/F items, but it is more difficult to determine the accuracy of the student response. Teacher ratings tend to become subjective unless scoring guidelines/rubrics are prepared beforehand.

PRO

1. True/False and matching items are relatively easy to write, sample a large amount of *subject* content, and are quickly and objectively scored.
2. True/False and matching items permit the widest sampling of content or objectives per unit of testing time.
3. Items require short periods of reading and response time, allowing you to cover more content.
4. Provides scoring efficiency and accuracy.
5. Produces highly reliable test scores.
6. Provides versatility in measuring all levels of cognitive ability.
7. Completion items are fairly easy to write, but less easy to grade because the correct answer may be subject to teacher interpretation. Grading guidelines should be constructed for each completion item prior to test administration.
8. Completion items can minimize guessing as compared to multiple-choice or true-false statements.
9. False items tend to discriminate more highly than true items. Therefore, use more false items than true items (but no more than 15% additional false items).

CON

1. True/false items do not discriminate between students of varying ability as well as other item types.
2. True/False, Matching and Completion tests frequently assess facts, encourage guessing, usually measure lower levels of cognitive learning, and often lead an instructor to favor testing of trivial knowledge.
3. True/False, Matching and Completion tests lack content and construct validity as they relate to "tasks taught."
4. True/False, Matching, Completion and Multiple Choice tests could be problematic if they do not cover all of the content that is taught (a validity problem).
5. True/False, Matching and Multiple-Choice test items are more precise in grading than are Completion test items.
6. True/False and Completion items are most often 'challengeable' and subject to interpretation because more than one answer may have to be considered correct if the item was not properly prepared.

7. Completion items have difficulty measuring learning objectives requiring more than simple recall of information.
8. Completion items require less time to construct, but more time to grade than multiple-choice or true-false items.

Criteria for Test Item Construction

Multiple-choice, true-false, matching and completion items represent test formats that require specific test developer skills. Included among these skills is both knowing the content to be tested and knowing the criteria and practices necessary to develop effective test items. Sevenair⁹⁰ developed a test item-writing checklist that includes:

1. Is the item clear and concise?
2. Did you use the active voice?
3. Did you avoid “ould” words
4. Is the difficulty level acceptable?
5. Does the stem pose a question or an incomplete thought?
6. If you used blanks, are they at the end of the stem?
7. Does the stem focus on a significant or important aspect?
8. Did you emphasize the NEGATIVES? (you shouldn’t)
9. Have you avoided keying the answer in the stem?
10. Are the distractors plausible?
11. Is there only one arguable correct response?
12. Are the foils (answers) homogeneous?
13. Did you avoid overlapping foils?
14. Are numerical foils in either ascending or descending order?

Glendale High School District⁹¹ provides a multiple-choice item guidelines checklist for its’ teachers. The twelve points include:

1. The stem clearly formulates a problem.
2. The stem contains only relevant information.
3. The vocabulary in the stem is at the correct level of difficulty for students.
4. The language or information in the stem is clear.
5. The stem is complete. It does not spill over into the response alternatives.
6. Verbal clues to the correct response are absent.
7. Response options are grammatically consistent with the stem.

⁹⁰ op.cit.

⁹¹ Becker, Marc S., 2002 *Multiple-choice Item Guidelines Checklist* Glendale Union High School District, Glendale, AZ.

8. Response options are parallel in length.
9. There is only one correct answer.
10. The negative is used sparingly or avoided.
11. “None of the above” or “all of the above” are avoided.
12. All distractors are plausible.

Comparable checklists for item writing have been made available in professional development activities conducted by Tempe Union, Apache Junction, Williams, Tucson Unified, Deer Valley and other districts.

PERFORMANCE TESTS (also called Authentic Assessment)

Custer et al⁹² has compiled an extensive text (68 pages) documenting issues and current practices in vocational education assessment. He notes that the terms alternative, authentic, and performance assessment conceptually and in practice tend to describe similar things. A copy of the six introductory pages to the monograph is included in Appendix H and is titled *Authentic Assessment—Basic Definitions and Perspectives*. A second, extensive background article titled *Authentic Assessment Tools* by Scott,⁹³ is also included in Appendix H.

Authentic assessment provides students with opportunities to:

1. connect and apply information,
2. obtain feedback or a running commentary on whether they are on course, and
3. provide a vehicle for self-assessment and reflection.

Authentic assessments are almost always framed in the form of learning experiences, ranging from simple to complex. They engage learners with “real time” information about the quality of their performance – while the performance is underway and not just after evaluation. These assessments connect the way schoolwork is assessed with the way knowledge and competence is judged in the workplace. As in the case of any assessment, alternative tests should be fair, valid, reliable, comparable, and generalizable. Cut-scores used for performance rating and grading purposes represent artificial points on a performance continuum and, therefore, represent an instructor challenge to be defensible.

Performance tests are most often used to provide students the opportunity to demonstrate/show their ability to apply content mastery in a specific knowledge area. The student is expected to perform correctly in a simulated, real-life situation.

⁹² Custer, Rodney L., Schell, John, McAlister, Brian D., Scott, John L., and Hoepfl, Marie “*Using Authentic Assessment in Vocational Education*” ERIC/ACVE IN 381. 2

⁹³ Scott, John (2002) *Authentic Assessment Tools* The University of Georgia. ERIC/ACVE, IN 381.2. 33-48

Students perform a task or series of tasks, rather than select an answer from a ready-made list as in true/false and multiple-choice test items. Many districts in Arizona are using performance-based tests (see p. 30-31).

Performance tests are based on the premise that students structure and apply information; therefore, they are engaged in “active” learning. The performance tasks present students with possibilities for applying an array of curriculum-related knowledge and skills. Both kinetic and artistic measures are/may be used. Tests have been developed for the assessment of vocational, managerial, administrative, leadership, communication, interpersonal and physical education skills in various simulated situations.

The quality of the student’s work is based on an agreed-upon set of criteria made known to students before they undertake the task(s). These are called scoring guides, rating sheets, and observation checklists. The ratings are used to provide students with a diagnosis of strengths/weaknesses and to contribute to an overall summary evaluation of the student’s abilities.

Indiana, Ohio, Oklahoma, Arkansas, New York, MAVCC, Wisconsin (WIDS), and the V-TECS consortium states (19) use various test item methodologies (including portfolios and scenarios) for performance assessment and student achievement reporting purposes.

PRO:

1. The use of performance tests is highly individual because it uses multiple measures and allows students to demonstrate significant tasks to solve real-world problems.
2. Tests focus on tasks that are meaningful to learners and are linked to school and non-school (industry) demands. Schoolwork is assessed in the way knowledge and competence is judged in the workplace.
3. Performance tests contain "items" referenced to specific application or demonstration of a skill (often in the psychomotor domain) or knowledge (cognitive). The curriculum drives the test administration, rather than a test requirement that disrupts the instructional sequence.
4. Performance tests require the specification and development of checklists, observation sheets, interview protocols, etc. for grading criteria. The criteria must be clear, high (ambitious), and defensible.
5. Performance tests are inherently instructional because they actively engage students in worthwhile learning activities. The activities encourage students to search out additional information or try different approaches. In some situations, the task(s) encourages students to work in teams.
6. The test is useful for measuring learning objectives, particularly in the psychomotor domain.
7. Usually provide a degree of test *validity* not possible with standard paper and pencil test items.

CON

1. Tests are difficult and time consuming to construct. The time needed to develop checklists, interview protocols, etc. for grading is exceptionally high.
2. Performance tests require individual rather than group assessment, such as on-site observation(s) by the instructor and/or a peer or industry representative using a checklist for each student's performance.
3. Performance tests cannot be as quickly graded as some other test forms.
4. In order to provide individualized observation and grading of student performance, teachers must allocate large blocks of time per class.
5. Performance assessment requires a greater expense of time, planning and thought from both students and teachers.
6. Generally provides low test and test scorer *reliability*.
7. Generally does not provide an objective measure of student achievement or ability (subject to bias on the part of the observer/grader).
8. Cut-scores used for performance rating and grading purposes could be challenged and need to be defensible.

Criteria for Writing Performance Test Items⁹⁴

1. Prepare items that elicit the type of behavior you want to measure.
2. Write items using a variety of measures (i.e. kinetic, artistic, etc.).
3. Clearly identify and explain the simulated situation to the student.
4. Make the simulated situation as "life-like" as possible.
5. Provide directions that clearly inform students of the type of response called for.
6. When appropriate, clearly state time and activity limitations in the directions.
7. Adequately train the observer(s) and scorer(s) to ensure that they are fair and consistent in scoring the appropriate behaviors.

Types of Performance Tests

- a. **Open-ended or extended response** exercises that require students to explore a topic orally or in writing following specific presentation criteria.
- b. **Extended tasks** that require sustained attention in a single work area and are carried out over several hours or longer. An example is painting a car in auto shop or producing a finished word-processed document within a certain time frame and to a pre-stated quality standard.

⁹⁴ op.cit. 19

- c. **Portfolios** that collect a variety of performance-based work (such as demonstrations and scenarios) to document improvements the student has made over time. They usually also contain the student's evaluation of the strengths and weaknesses of several pieces included in the portfolio. They may also contain presentations of previously evaluated materials to show historical improvements made in the portfolio item(s).

PORTFOLIO ASSESSMENT:

Portfolios initially were a way for artists, graphic designers, and others to show evidence of their creative work; it was a repository of work samples. Today, portfolios include contents such as examples, reviews, demonstrations and ratings of students' vocational skills, and evidence of academic achievements. Portfolio assessment is particularly responsive to integrated curriculum. Tucson Unified District provides District guidelines for portfolio assessment. Other districts surveyed are/may be using them, but do not require use of any district guidelines. Portfolio use is primarily a teacher assessment option in most schools surveyed.

PRO

1. The portfolio provides multiple, tangible evidence of student accomplishment in a format transferable to the job search. It reflects the breadth of study in the curriculum and the quality of work that students are expected and able to produce.
2. Portfolio assessment responds to integrated curriculum and accountability standards.
3. Portfolios provide flexibility because the content quantity may be either comprehensive or selectively-based.
4. Portfolio criteria are tools to help students see their own learning gaps, set goals for future experiences, manage and monitor their learning, and document their changes/improvement over a period of time. Assembling a portfolio encourages pride in ones' work and improved self-concept.
5. Portfolio assessment is not teacher driven and supports student-centered classrooms; it is a shared responsibility among students, teachers, parents, and employers.
6. Portfolios accommodate the diverse learning patterns of all students and enables each of them to realize and experience success.
7. Portfolios motivate students because it gives them some control over what and how they learn and how their performance will be assessed.

CON:

1. Reliability is a concern when the portfolios contain different pieces and have diverse purposes.
2. Lack of standardization in the way portfolio entries are produced and the amount of assistance students receive presents a problem.
3. Portfolios may not reflect sustainable levels of performance under normal conditions.

4. Teachers may not be equipped to conduct effective portfolio assessments without staff development training and time to collaborate with other instructors to develop portfolio-rating criteria.

Criteria for Portfolios:

Portfolios should include:

1. A thoughtful student developed introduction to the portfolio.
2. Student developed reflection papers behind each major assignment of the portfolio.
3. Student self-assessment of their own work through scoring rubrics for portfolio entries.

Teacher developed materials to support portfolio assessment should include:

1. Established models, standards, and criteria that assist students in selecting their best work to include in the portfolio.
2. Rubrics for each type of portfolio entry. (See templates below for Tech Prep, TUHSD and #1-12 “Rubrics Madness”)
3. Established procedures for student oral presentations of their portfolios to significant others (peers, teachers, parents) and presentation criteria.

PERFORMANCE-BASED, ESSAY, AND SUBJECTIVE TESTS

Subjective tests permit students to organize and present original answers to specific questions. Included in this category of tests are short-answer essay, extended-response essay, problem-solving and performance-based test items. Ebel⁹⁵ has written extensively on essay tests and proposes that they are especially appropriate when:

- The group to be tested is small and the test is not to be reused.
- You wish to encourage and reward the development of student writing skills.
- You are more interested in exploring the student’s attitudes rather than in measuring his/her achievement.
- You are more confident of your ability as a critical and fair reader than as an imaginative writer of good objective test items.

The Office of Instructional Resources at the University of Illinois Urbana-Champaign campus has developed a comprehensive resource booklet of the Ebel text referenced above. The resource booklet, titled “Improving Your Test Questions” (<http://www.oir.uiuc.edu/dme/esams/ITQ/html>) includes sections on all test types with pro/con statements, sample questions, guidelines for developing the test type, and suggestions for scoring.⁹⁶

⁹⁵ Ebel, Robert L. Measuring educational achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965, Chapters 4-6, and Ebel, Robert L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972, Chapters 5-8.

⁹⁶ University of Illinois Urbana-Champaign “Improving Your Test Questions” Urbana, IL.
<http://www.oir.uiuc.edu/dme/exams/ITQ.html>

Essay tests are most commonly used to allow the student freedom of response to a given problem. They require students to recall information, as well as organize and express ideas clearly and, usually, within a certain time frame. Scoring criteria (rubrics) encourage use of appropriate written style and language. Most authors recommend developing essay questions *before* presenting the material, including the scoring rubric. This, in turn, directs instruction and teachers are more likely to teach the criteria and content completely.

Scoring models for essay items are 1) analytical scoring and 2) global quality. In the analytical scoring model, each answer is compared to an ideal answer and points are assigned for the inclusion of necessary elements. In the global quality model, each answer is read and assigned a score based either on the total quality of the response or on the total quality of the response *relative to other students' answers*.

Open-ended questions stressing strategies and problem solving may be measuring general ability (the old I.Q. of intelligence testing) rather than mastery of content. It is important to note that students do not have IQ's; they "earn" IQ scores based on a particular test score interpretation.

PRO:

1. Essay tests can be used to assess a student's attainment of specific instructional objectives at higher levels of the cognitive domain (i.e. analysis, synthesis, and evaluation).
2. Essay tests allow students to express themselves in their own words and to organize concepts in ways that are meaningful/useful to them.
3. Essay tests can be designed around realistic, work-related scenarios.
4. Essay items are easier and less time consuming to construct than are most other item types.
5. Essay and other free form exams can diagnose more than multiple-choice items when direct improvement of learning through individual diagnosis and instructional feedback is desired.
6. Essay or written performance questions can show higher degrees of student understanding and functioning than multiple-choice questions designed at the very highest levels of complexity.
7. Scoring may be analytical (each answer is compared to an ideal answer) or global (total quality of the response or response compared to other students).

CON

1. Essay tests have a major disadvantage in that the level of objectivity associated with rating student responses may vary, and test graders may be inconsistent. (Inter-rater consistency can be improved by providing grading rubrics.).
2. Subject to bias on the part of the grader.
3. Generally provides low test and test-scorer reliability.

4. Essay tests, by their nature, typically sample only a small segment of the domain (may emphasize one topic over another with which a student may or may not be familiar).
5. Cannot measure a large amount of content or objectives.
6. Compared to multiple-choice tests of similar length, written examinations are time consuming and less objective than multiple-choice tests when being graded.
7. Requires an extensive amount of instructor's time to read and grade.
8. Generally does not provide an objective measure of student achievement or ability.

Criteria for Essay Tests:

1. Develop realistic, work-related scenarios.
2. Specify related instructional objectives.
3. Identify primary purpose of the essay (i.e. recall information; synthesize, etc.)
4. If using open-ended questions, stress strategies and problem-solving approaches.
5. Establish a certain time frame for completion of the essay.
6. Establish minimum/maximum length requirements, if desired.
7. Develop scoring criteria (grading rubrics) to minimally include:
 - a. higher levels of the cognitive domain (i.e. synthesis and evaluation),
 - b. ability to compose an answer and present it in a logical manner,
 - c. evidence of ability to organize and express ideas clearly,
 - d. demonstration of appropriate writing skills,
 - e. evidence of "correct" problem solution or recommended strategy.
8. Provide time to review students' scores with them and to re-direct instruction.

PROBLEM SOLVING TESTS (also called SCENARIOS or COMPUTATIONAL EXAM QUESTIONS)

Problem solving items present a problem situation or task and require the student to demonstrate work procedures and a correct solution. Variations include only presenting a correct solution. This test type is classified as subjective type because of the procedures used to score the item responses.

Charles Losh, a proponent for standards-based instruction and scenarios as an instructional technique to integrate academic and vocational content states: "derived from the real world, skill standards provide a natural content base for contextual instruction. Standards-based scenarios provide a rich context for the integration of academic and vocational education."⁹⁷ He notes that once contextual skills are identified they enhance dialog between vocational and academic instructors. His report includes appendices containing sample skill standards, model scenarios, a checklist for instructional/assessment criteria, a sample scenario planner, and a blank master.

Information on using skill standards for curriculum development may also be found in the ERIC database using the following descriptors: Competence, Criterion Referenced Tests,

⁹⁷ Losh, Charles L. (2002) *Using Skill Standards for Vocational-Technical Education Curriculum Development Information Series No. 383* ERIC/ACVE, Washington DC and (2000) *The Linkage System: Linking Academic Content Standards and Occupational Skill Standards* (Ver 1.2) V-TECS, Southern Association of Colleges and Schools. 8 <http://www.v-tecs.org>

*Curriculum Development, *Job Skills, *National Standards, *Performance Based Assessment, *State Standards, and Vocational Education, and the identifier *Scenarios. Asterisks indicate descriptors that are particularly relevant to scenarios.

PRO:

1. Minimize guessing by requiring students to provide an original response rather than to select from several alternatives.
2. Easier to construct than multiple-choice or matching items.
3. Most appropriately used to measure learning objectives which focus on the ability to apply skills or knowledge *in the solution of problems*.
4. Can measure an extensive amount of content or objectives.

CON:

1. Generally provide low test and test scorer reliability.
2. Requires large amount of instructor time to read and grade.
3. Not an objective measure of student achievement or ability (subject to bias on the part of the grader when partial credit is given).

Criteria for Problem Solving and Scenario Test Items:

University of Illinois guidelines⁹⁸ for writing problem-solving test items include:

1. Clearly identify and explain the problem.
2. Provide directions that clearly inform the student of the type of response called for.
3. State in the directions whether or not the student must show his/her work procedures for full or partial credit.
4. Clearly separate item parts and indicate their point value (scoring rubrics).
5. Use figures, conditions and situations that create a realistic problem.
6. Ask questions that elicit responses on which experts could agree that one solution and one or more work procedures are better than others.
7. Work through each problem before classroom administration to double-check accuracy.

Performance-based Assessment: Performance-based assessments are generally thought of as demonstrations and are widely used in career/technical programs. Teachers generally use checklists to verify process steps. Some teachers also use checklists with rating scales to assess quality of the task(s) performed. Checklists are appropriate for scoring many activities; however, there are performance activities that require more precise measures.

Scott⁹⁹ developed an extensive list to assist teachers in identifying potential instructional activities for performance-based assessment, presented in Table VIII on the next page. To assess the adequacy of performance activities listed in the table, more detailed information than a process checklist is necessary. This detail is accomplished by using a rubric which is defined as a “scale and list of characteristics describing performance for each of the points on

⁹⁸ op.cit. 18

⁹⁹ Scott, John L., “Alternative Assessment Tools” in Using Authentic Assessment in Vocational Education ERIC/ACVE IN 381. 29

the scale.”¹⁰⁰ Rubrics communicate and assign performance criteria to measure complex activities/tasks and provide an objective basis for scoring acceptable performance. Rubrics provide both the detail and the means of establishing cut scores/values for meaningful assessment of these activities.

Table VIII: Authentic Assessment Tools/Performance Activities¹⁰¹

Graphic Organizers and Concept Mapping		
Concept Maps	Correlation/scatter diagrams	Event chains
Data tables		Histograms
Cause/Effect Diagrams	Idea webs/graphic organizers	PMI strategy reports
Graphs	Geographic maps	Mrs. Potter's questions
Run control charts	Time Lines	Connecting elephants
Flowcharts	Venn Diagrams	Big Idea generation
Pareto diagrams		Ranking ladders
		Mind maps

Performance Products		
Business letters	Vitas/Resumes	Pamphlets
Autobiographies	Inventions	Observation reports
Editorials	Lab reports	Research reports
Displays	Information-seeking letters	Posters
Drawings/illustrations	Management plans	Workplace scrapbooks
Experiments	Math problems	Grant applications
Essays	Geometry problems	Team reports
Surveys	Models	Career plans
Storyboard Reports	Writing samples	Video yearbooks
Job Applications	Job searches	Training plans
Book reviews	Cartoons or comics	Exhibits

¹⁰⁰ Marzano, Pickering, and McTighe in Scott, John L., in Using Authentic Assessment in Vocational Education ERIC/ACVE IN 381. 10

¹⁰¹ op.cit. 39

Performance Products		
Bulletins	Collages	Ballads
Critiques	Consumer reports	Announcements
Crossword puzzles	Handbooks	Biographies
Designs	Booklets	Questionnaires
Requisitions	Home projects	Technical repairs

Live Performances and Presentations		
Interviews	Games/quiz bowls	Commercials
Issues/controversy	Student-led conferences	Demonstrations
Workplace skits	Story time/anecdotes	Newscasts
Slide shows/video	Prepared and extemporaneous speeches	Plays-TV/radio broadcasts
Human graphs		
Announcements		

Performance activities (e.g. demonstrations, team reports, essays, workplace skits, scenarios, technical repairs) provide a unique opportunity for students to demonstrate higher order skills including application, analysis, synthesis, and evaluation. As teachers develop more appropriate learner expectation statements (standards, learner objectives, scoring rubrics), they also assist students in directing their own learning and in meeting and exceeding the standards.

RUBRICS and CHECKLISTS FOR PERFORMANCE ASSESSMENT

Rubrics, also called scoring guides, use *pre-established performance criteria* to reduce teacher/scorer subjectivity and allow more objective evaluation of student performance. Rubrics are most often used with performance assessments and may be general or task specific (however, it is possible for a rubric to contain both elements (general and task specific). Scoring rubrics are one of the most common methods for assessment. In 1993, Marzano, Pickering, and McTighe defined rubrics as “a fixed scale and list of characteristics describing performance for each of the point on the scale.”¹⁰² Rubrics are tools to clarify, communicate, and assess performance based on criteria that are often complex and subjective. The rubric encourages objectivity in scoring practices.

The scoring sheet (rubric) lists criteria in gradations of quality to assess a product. No single form is the “right one” to use, as rubrics are highly dependent on the activity to be evaluated. Rubrics are not dependent on grade level or content, but rather on the purpose of the assessment. Foremost, a rubric involves a judgment of quality of work in whole or in part. Thus, each criterion within any rubric may be scored on a *different descriptive scale*.

Mertler¹⁰³ identified three scoring instruments for performance assessments including checklists, rating scales and rubrics with rubrics divided into two types: analytic and holistic. Checklists are utilized in noting elements within a performance or demonstration activity (process); the emphasis in a checklist is in noting *sequences* within an activity and the degree to which each step is completed successfully (process and level of acceptability).

Rating scales—as opposed to checklists—combine task sequences with point values to determine acceptable performance. In simple form, rubrics are checklists requiring a yes or no response. More complex rubrics differentiate levels of performance and contain indicators describing student performance that meets or exceeds the standard.

Holistic rubrics score the overall process/product as a whole; component parts are not judged separately. A variety of performance requirements are considered when constructing the rubric, but each separate requirement is *not* given a separate rating scale. A single descriptive scoring scheme is used for the overall performance. Holistic rubrics provide information on overall performance and *do not* give specific information for improvement.

Analytic rubrics involve scoring separate, individual parts of the product or performance first, followed by summing the individual part scores to obtain a total score. Each criterion is considered separately as descriptions of the different score levels are developed and each evaluation factor (description) receives a score. After all parts are scored, a composite score is developed. Brookhart¹⁰⁴ concluded that scores should represent meaningful distinctions

¹⁰² Marzano, Pickering, and McTighe in Scott, John L., in Using Authentic Assessment in Vocational Education ERIC/ACVE IN 381. 10

¹⁰³ Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). Available online: <http://ericae.net/pare/getvn.asp?v=7&n=25>

¹⁰⁴ Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27. No. 1) Washington, DC: The George Washington University, Graduate School of Education and Human Development.

between score categories, thus providing students an opportunity to objectively compare their work to the criterion and establish learning goals for additional skill(s) development. Analytic rubrics usually include two or more separate scales.

Mertler¹⁰⁵ provides a template for holistic and analytic rubrics, sample grades and categories, rubric design procedures, and a sample scoring rubric for a performance task. His article is included in Appendix H and is titled *Designing Scoring Rubrics for Your Classroom*. A second article, written by Moskal¹⁰⁶ titled *Scoring Rubrics: What, When and How?* is also the Appendix H. Other resources include the ERIC Clearinghouse on Assessment and Evaluation (ERIC/AE) which links to several web sites, including *Scoring Rubrics – Definitions & Constructions* available on line @ http://ericae.net/faqs/rubrics/scoring_rubrics.htm Chicago Public Schools has developed a Performa Online Resource for RUBRICS (Rubric Bank) @ http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Rubric_Bank/rubric_bank.html A link is also available through <http://www.napehq.org/>

PRO

1. Uses pre-established performance criteria.
2. Focuses attention on important outcomes with an assigned value for each.
3. Lowers student anxiety about what is expected of them.
4. Provides an objective scoring basis and reduces teacher/scorer subjectivity.
5. Objectively evaluates student performance and demystifies scoring practices.
6. Ensures equity in scoring practices (all work is judged by the same standard).
7. Promotes formative (process) as well as summative (product) evaluation.
8. Improves rater reliability (improves consistency).
9. May target general or task goals/performance.
10. Provides discussion information for teachers to use with students.

CON

1. Time consuming task to develop descriptions, criteria, and levels of attainment.
2. Teachers tempted to adopt rubrics from other sources (which may not be appropriate for their instructional goals, task descriptions and/or proficiency levels).
3. Requires additional time to counsel students and provide improvement feedback.

Criteria for Rubrics (See also detailed guidelines # 1-4 following Table VIII):

1. Identify qualities student must demonstrate.
2. Develop the performance criteria.
3. Establish a descriptive rating scale .
4. Identify levels of proficiency.

¹⁰⁵ *Op.cit*, 2-3

¹⁰⁶ Moskal, Barbara M., (2000) *Scoring Rubrics: What, When and How?* Practical Assessment, Research & Evaluation, 7(3) ERIC Clearinghouse on Assessment and Evaluation. Available online: <http://ericae.net/pare/getvn.asp?v=7&n=3>

Write descriptions (characteristics) of performance proficiency levels. Rubric Guidelines and Templates:

The following section contains four types of information: guidelines for developing rubrics, samples of two Arizona-based rubrics, a sample of a testing service rubric, and abbreviated samples of twelve rubric templates.

At the most recent ADE statewide conference for vocational educators, a four-hour rubrics workshop was conducted (7/22/02) by Ms. Judy Balogh (Arizona State University), Michelle Crary (Desert Vista High School), Chris Libette Garcia (Metro Tech High School) Sue Crumrine (Winslow High School) and Nanette Gillispie (Peoria School District). The workshop was titled “Rubric Madness!¹⁰⁷ Student and Program Assessment Made Easy for Business Education.” Workshop materials included a reference notebook detailing benefits, elements, criteria, construction hints, converting rubrics to grades, and samples of rubrics. Abbreviated samples of the rubrics templates are presented in some the materials that follow.

Guideline 1: Designing rubrics: **Steps To Developing Rubrics**

1. Determine which concepts, skills or performance standards you are assessing.
2. List the concepts and rewrite them into statements that reflect both cognitive and performance components.
3. Identify the most important concepts or skills being assessed in the task.
4. Determine the number of points to be used for the rubric. Determine the description for each score based on the importance of each element.
6. Compare student work to the rubric looking for gaps in required elements.
7. Revise the rubric description to include any items not originally captured.
8. Rethink and readjust the scale to make sure the points differentiate enough.
9. Write descriptions (characteristics) of performance proficiency levels.

Guideline 2: Designing rubrics: **Rubric Design Protocol That Works...if all else fails!**

1. Gather student work samples.
2. Sort samples into 3 or 4 groups.
3. Record your own descriptive statements.
4. Categorize statements into critical performance elements.
5. Write an operational definition of each element.
6. Select the “best match” of student work per each level of performance (i.e., exemplary, proficient, basic, novice).
7. Repeat steps...refining your rubric’s elements, descriptors, and indicators.
8. Store your rubrics and students work examples at each level to use for instructional and communication purposes.

Guideline 3: Designing rubrics: **Why Rubrics Are Important**

¹⁰⁷ Balogh, Judy; Crary, Michelle and Libette Garcia, Chris. *Rubric Madness! Student and Program Assessment Made Easy For Business Education*” Paper presented at the Arizona State Annual Vocational Conference, July 22, 2002. Tucson, Arizona. Copies available from Ms. Judy Balogh (Arizona State University and/or Dr. Janet Gandy (Arizona Department of Education

Guidelines for Rubric Construction

Why Rubrics Are Important

A good scoring rubric will:

- ☐ Help teachers define excellence and plan to help students achieve it.
- ☐ Communicate to students what constitutes excellence and how evaluate their own work.
- ☐ Communicate goals and results to parents and others.
- ☐ Help teachers or other raters be accurate, unbiased, and consistent in scoring.
- ☐ Document the procedures used in making important judgments about students.

Steps in Rubric Development

1. Determine learning outcomes
2. Keep it short and simple (4-5 items; use brief statements or phrases)
3. Each rubric item should fit on one sheet or paper
4. Focus on how students develop and express their learning
5. Evaluate only measurable criteria
6. Ideally, the entire rubric should fit on one sheet of paper
7. Reevaluate the rubric (Did it work? Was it sufficiently detailed?)

Terms To Use in Measuring Range/Scoring Levels

(Numeric scale ranging from 1 to 5)

Needs Improvement...Satisfactory...Good...Exemplary
Beginning...Developing...Accomplished...Exemplary
Needs work...Good...Excellent
Novice...Apprentice...Proficient...distinguished

Concept Words that Convey various Degrees of Performance

(After you write your first paragraph of the highest level, circle the words in that paragraph that can vary. These words will be the ones that you will change as you write the less than top level performances.)

Depth...Breadth...Quality...Scope...Extent...Complexity...Degrees...Accuracy

Presence to absence

Complete to incomplete

Many to some to none

Major to minor

Consistent to inconsistent

Frequency: always to generally to sometimes to rarely

Guideline 4: Designing rubrics: **Adjectives and Adverbs for Rubric Construction**

Use this chart to “jump-start” your rubric construction. Your mission is to clarify these terms by adding specific requirements. Avoid vague terms by qualifying your descriptors and defining exactly what you want from the students.

Not Meeting Expectations	Progressing	Proficient	Exemplary
None Never Incomplete Inadequate Unsatisfactory Unclear Rarely clear ...to an unacceptable level Includes no elements of... Improper Unclear Inappropriate Lacks enough of... Inconsequential, Unimportant Unnecessary Illogical Random	Fewer than ____ Seldom, rarely Less than complete Less than adequate Minimal Vague Sometimes unclear/inaccurate ...to a minimal level Includes few elements of... Sometimes improper Somewhat unclear Limited Minimal amount of... Somewhat relevant Somewhat useful Somewhat reasonable Somewhat instinctive	More than ____ Sometimes, often Somewhat complete Adequate Satisfactory Understandable Often clear, often accurate ...to an acceptable level Includes most elements of... Somewhat proper Some degree of clarity Somewhat appropriate Adequate number of... Important Essential Reasonable Somewhat intuitive	All Always Complete Superior Maximum Articulate Clear, accurate ...to the highest level Includes all elements of... Clear Proper Appropriate All Necessary... Significant Critical, crucial Logical, rational Intuitive

Sample 1: **Confirming rubrics:** EVALUATING EXISTING RUBRICS

- ☐ Does the rubric relate to the outcomes being measured?
- ☐ Does the rubric cover important dimensions of student performance?
- ☐ Do the criteria reflect current conceptions of “excellence” in the field?
- ☐ Are the categories or scales well-defined?
- ☐ Is there a clear basis for assigning scores at each scale point?
- ☐ Can the rubric be applied consistently by different scorers?
- ☐ Can the rubric be understood by students and parents?
- ☐ Is the rubric developmentally appropriate?
- ☐ Can the rubric be applied to a variety of tasks?
- ☐ Is the rubric fair and free from bias?

Sample 2: Confirming rubrics **A Rubric Checklist**

Use the checklist to review your own rubrics to determine adequacy.

Performance Criteria: *What is being evaluated?*

- Are the performance criteria linked to standards?
- Are there a manageable number of performance criteria? (Usually 3-5 is adequate.)
- Are the performance criteria measurable/teachable?
- Do performance criteria match the objectives?

Comments
By _____
Title of Rubric _____

Key Question: Are the performance criteria clearly stated with a meaningful label?

Scales and Levels of Proficiency: *Degrees of quality?*

- Is there an even number of levels, i.e., 4 to avoid middle scoring?
- Does the highest level represent exemplary performance?

Key Question: Is each level on the scale meaningful and non-judgmental?

Descriptions: *What would success look like?*

- Are they written in student language? Clear and understandable?
- Are they positively stated?
- Are the differences in descriptions observable? Clearly stated expectations?
- Is there a progression of differences among the descriptions?

Key Question: Does the rubric have these elements: Performance Criteria, Scale, Levels of Proficiency, and Descriptions?

Key Question: Is the rubric manageable and practical to use by students and teachers? Have others (e.g., peers and students) reviewed this rubric?

Rubric templates: As noted earlier, there is no one single “right” rubric to use. Instead, each rubric can be unique to the instructional program, instructional content, and desired performance levels. Rubrics may guide individual performance and/or be used to assess program effectiveness. Experience has shown that rubrics using a four to six point scoring range are quite effective. Three sample rubrics to use with vocational students and one sample rubric to use with special needs students are presented and include:

1. Arizona Tech Prep which uses a five-point scale (based on 0-3 scaling) and has extensive performance descriptions.
2. Tucson Unified School District, which uses a four-point scale that is less extensive, but still contains adequate descriptors.
3. Rubrics from the CTB/McGraw-Hill’s Tests of Adult Basic Education (TABE) “Adventures in Writing section.”

4. South Carolina Rubric for Special Needs Students (PACT-Alt) that uses scoring dimensions and four attainment levels of proficiency.

Following these samples, readers will find abbreviated versions of twelve sample templates from the “Rubrics Madness” workshop materials described earlier.¹⁰⁸

¹⁰⁸ Balogh, Judy; Crary, Michelle and Libette Garcia, Chris. *Rubric Maness! Student and Program Assessment Made Easy For Business Education*” Paper presented at the Arizona State Annual Vocational Conference, July 22, 2002. Tucson, Arizona. Copies available from Ms. Judy Balogh (Arizona State University and/or Dr. Janet Gandy (Arizona Department of Education

Sample 1: **Arizona Tech Prep Rubric:** The following sample rubric is from the Arizona Tech Prep web site <http://www.aztechprep.org/vocationalprograms>. The rubric uses a five-point attainment scale (based on 0-3 ratings) and using extensive descriptors. The rubric is used by many districts because it can be adapted for any/all Arizona CTE programs. This sample is a variation of sample templates # 1,2, and 12 that follow.

RUBRIC FOR ASSESSMENT OF INDIVIDUAL COMPETENCY ATTAINMENT

AGRICULTURE - Agriscience LEVEL III

School

Teacher

LEVEL OF ATTAINMENT (LOCAL DISTRICT PERCENTAGES MAY BE DIFFERENT THAN SAMPLE)				
3	2	1	0	0
90% +	80% +	70% +	60% +	Less than 60%
MASTERED	EXCEEDED	ATTAINED	<i>Approaching Attainment</i>	<i>Unattained</i>
Student presents a clear, specific understanding of the competency. All notes, assignments, test, workplace records and labs required are completed on time, are extremely well organized and questions are answered accurately. High interest and excitement have lead the student to reach far beyond the requirements. Student has read related materials and has used many sources of information for reports and or experiments. The student has used his/her new knowledge when participating in all oral discussions, assignments and written work. Student makes connections between classroom and workplace. The students' notes, tests, labs, workplace records, debates, CTSO participation, and assignments are of the highest level of achievement above 90%.	Student presents a clear, specific understanding of the competency. High interest and excitement leads the student to an investigation that reaches beyond requirements. All notes, assignments, tests, workplace records and labs required are completed on time, are very well organized and questions are answered accurately. The student has used more resources than required and demonstrates new knowledge both orally and in written work and uses this knowledge in his/her assignments and oral participation. New knowledge is evident when student shows connections between classroom and workplace relationships. Student notes, tests, labs, work place records, CTSO participation, debates and assignments are clearly organized, carefully done, and often go beyond teacher expectations. All tests are beyond the standard level of achievement between 80% to 89%.	Student meets assignment expectations. The student demonstrates new knowledge learned in oral participation and or written tasks. The work is well organized and complete. The student understood the assignments. He/she used the resources required and organized information in all notes, assignments, tests, workplace records, debates and labs. All notes, assignments and labs are complete, carefully done and the student meets just above the minimum requirements and expectations. All tests, workplace records, CTSO participation, assignments and labs meet the standard level of achievement between 70% to 79%.	Student knowledge of the topic is understood, but at minimum level of competency. The assignments, notes and labs are occasionally incomplete and could be organized better. Some resources have been used, but it is not clear what the student understood. Some of the information included by the student was not important to the topic. Student does most of what is required, but nothing more. Some of the work may not be finished. Tasks are not carefully done and the information from the resources is not used. Tests, labs, notes, CTSO participation, and work-based learning results are at a level of achievement between 60% to 69%.	Student knowledge of the subject is not shown. Steps through the process were not followed. Notes, tests, assignments, work-based learning and labs lack neatness, organization, detail and evidence of new knowledge. Work does not meet requirements. Parts are missing. Participation is weak, or student is often not participating. Labs, tests, CTSO participation, and assignments are poorly done and fall well behind the standard level of achievement. Overall, the student has failed to grasp new concepts covered in the competency. The level of achievement is below 60%.

Sample Arizona Rubric 2: **Tucson Unified School District**: As a point of comparison, the following rubric is used in the Tucson Unified School District¹⁰⁹ for competency attainment scoring. Readers will note that the descriptions are succinct and easily followed. This sample demonstrates a four-point scale using less descriptive, but still adequate, criteria than those demonstrated in the Tech Prep sample above. In addition, each teacher needs to equate the scoring scale or series of scales to reflect grading practices and 80% or more competency attainment for the course/program. This sample is comparable to sample template # 4 below. Deer Valley Unified District and several others also use District-developed rubrics.

Criteria for Marking Competency Attainment
(Tucson Unified School District, Tucson, AZ)

CTE Skills	State-designated competencies for a program area
Academic Standards	State-designated standards for language arts, math, and science

Scoring Scale:

Score Point	Score Point Description
4	<p>The “4” response reflects a thorough knowledge and understanding of the skill.</p> <ul style="list-style-type: none"> • The purpose of the assignment is fully achieved. • There is a substantial, accurate, and appropriate application on content knowledge. • The response reflects an ably reasoned argument in relation to the assigned topic.
3	<p>The “3” response reflects an adequate knowledge and understanding of the skill.</p> <ul style="list-style-type: none"> • The purpose of the assignment is largely achieved. • There is a generally accurate and appropriate application of content knowledge. • The response reflects an adequately reasoned argument in relation to the assigned topic.
2	<p>The “2” response reflects a limited knowledge and understanding of the skill.</p> <ul style="list-style-type: none"> • The purpose of the assignment(s) is partially achieved. • There is a limited, possibly inaccurate or inappropriate application of content knowledge. • The response reflects a limited, poorly reasoned argument in relation to the assigned topic
1	<p>The “1” response reflects a weak knowledge and understanding of the skill.</p> <ul style="list-style-type: none"> • The purpose of the assignment is not achieved. • There is little or no appropriate or accurate application of content knowledge. • The response reflects little or no reasoning in relation to the assigned topic.

¹⁰⁹ Prather, Kathy, Tucson Unified School District, Tucson, AZ.

Sample Testing Service Rubric 3: Test of Adult Basic Skills (TABE) “Adventures in Writing”

TABE is a commercial basic skills assessment published by CTB/McGraw-Hill widely used in secondary schools. The test is normed on students in grades 9 and higher and tests ability in Reading, Language, Mathematics Computation, Applied Mathematics, and Spelling. The sample rubric is included to show a seven-point scale/rubric that could be used to score *writing skills* on essay test forms. A second rubric to score *technical content* would also need to be used in conjunction with the writing skills rubric.

Rating	Descriptions
R	Response cannot be scored because of quantity of words produced, or Illegibility of handwriting.
0	No response or merely copies the prompt.
1	Response consists of only isolated words, phrases, or dependent clauses with no complete sentences.
2	Responses do not focus on a single idea or event. May be only two or three disjointed sentences.
3	Responses contain an identifiable story line, with a beginning, a middle, and (usually) an ending.
4	Responses contain a clear series of events or ideas. Word choice, sentence structure, and organization may be simple in part.
5	Responses are vivid and precise in vocabulary. Sentence structure is fluent and marked by use of accurate and varied transitional signals.

Sample Rubric 4: **South Carolina Rubric for Special Needs Students (PACT-Alt)**

The following rubric is specifically designed for use with disabled students (grades 3-8) who need accommodation, who participate in Special Education programs, and who require alternative assessments to measure attainment on state academic standards. The rubric is used in South Carolina and several other states participating in the Mid-South Regional Resource Center at the University of Kentucky and the Inclusive Large Scale Standards & Assessment (ILSSA) project. With modification, the scoring guide could be adapted for use with IVEP students in Arizona vocational programs. Readers should note that this guide references four grading periods rather than one grading period.

PACT-Alt Scoring Guide

Student Performance

Provides information on student's progress on the functional targeted skill within the context of the SC Curriculum Standards.

Scoring Dimensions	Below Basic	Basic	Proficient	Advanced
Student Progress within standards based activities	Data recorded in all 4 periods Progress on functional targeted skill not evidenced Increased complexity not present or clear	Data recorded in all 4 periods Progress on functional targeted skill evidenced in the 2nd, 3rd, and 4th periods Increased complexity not present or clear.	Data recorded in all 4 periods Progress on functional targeted skill evidenced in the 2nd, 3rd, and 4th periods Increased complexity evidenced in 2 of the last 3 periods (2nd, 3rd, 4th periods)	Data recorded in all 4 periods Progress evidenced on functional targeted skill in the 2nd, 3rd, and 4th periods Increased complexity evidenced in the 2nd, 3rd, and 4th periods

Program Supports

Provides information on effective practice and program supports for student performance on the targeted functional skill.

Scoring Dimensions	1	2	3
Standards Based Activities	There is little or no evidence of opportunity for the student to perform functional targeted skill within the context of age-appropriate standards based activities	There is evidence of opportunity for the student to perform functional targeted skill within the context of age-appropriate standards based activities.	There is evidence of opportunity for the student to perform functional targeted skill within the context of a variety of age-appropriate standards based activities.
Opportunity For Student Self-determination within Standards Based Activities	There is little or no evidence of opportunity for the student to make choices.	There is evidence of limited opportunity for the student to make choices that impact student learning.	There is evidence of consistent opportunity for the student to make choices that impact student learning.
Opportunity for Standards based Instruction within Multiple Settings	There is no evidence that the student receives instruction and has the opportunity to perform the functional targeted skill in settings other than specialized environments.	There is evidence that the student receives instruction and has the opportunity to perform the functional targeted skill in a variety of settings.	There is evidence that the student receives instruction and has the opportunity to perform the functional targeted skill in a variety of settings, at least one of which must be with non-disabled peers or in the community.

Sample Templates: #1-12: The following twelve *abbreviated sample rubric templates* are from “Rubric Madness! Student and Program Assessment Made Easy For Business Education.”¹¹⁰ Depending on the tasks, scoring criteria, activity to be evaluated, and purpose of the assessment, an appropriate rubric format can be selected and developed. Any rubric that involves a judgment of quality of work, in whole or in part, may be the “correct” rubric to use.

Rubric Template #1

Rubric Name: _____ Date _____

Comments: _____

	<u>Superior Performance Indicator</u>	<u>Sufficient Performance Indicator</u>	<u>Limited Performance Indicator</u>	<u>Minimum Performance Indicator</u>
First Objective	Performance Indicator	Performance Indicator	Performance Indicator	Performance Indicator
Second Objective	Performance Indicator	Performance Indicator	Performance Indicator	Performance Indicator

Rubric Template #2:

Rubric Name:

Objective or Performance	Beginning – 1	Developing – 2	Accomplished – 3	Exemplary – 4	Score

¹¹⁰ Balogh, Judy; Crary, Michelle and Libette-Garcia, Chris. *Rubric Madness! Student and Program Assessment Made Easy For Business Education*” Paper presented at the Arizona Annual State Vocational Conference, July 22, 2002. Tucson, Arizona. Copies available from Ms. Judy Balogh (Arizona State University and/or Dr. Janet Gandy (Arizona Department of Education, Career and Technical Education Division).

Rubric Template #3 (Name of Rubric)

Directions: Circle the number on the right of each category to indicate overall quality for the performance category

Next, mark "+" for strengths on the blanks to the left of each quality indicator within categories.
1=Unskilled; 2=Poorly Skilled; 3=Moderately Skilled; 4=Skilled; 5=Highly Skilled

Category #1: _____	1	2	3	4	5

Category #2: _____	1	2	3	4	5

Rubric Template #4

Name of Rubric: _____

Categories	Criteria	Quality of Performance (check one)
		<input type="checkbox"/> Lack of evidence <input type="checkbox"/> Clear evidence
		<input type="checkbox"/> Lack of evidence <input type="checkbox"/> Clear evidence

Rubric Template #5

Name of Rubric: _____

<u>Goals of Assignment:</u> 1. 2. 3. 4. 5.	<u>What to do for an B:</u> 1. 2. 3. 4. 5.
<u>What to do for an A:</u> 1. 2. 3. 4. 5.	<u>A student will be marked N for the following reasons :</u> 1. 2. 3. 4. 5.

Note: Students who receive an **N** will be expected to continue work on assignment until they have achieved either an **A** or **B**.

Rubric Template #6

Name _____ Date _____

Scoring Guide for _____

Categories with descriptions:	Score Value						
	6	5	4	3	2	1	Off Task

Rubric Template #7

Your Name: _____ Group Topic: _____

Group Members: _____

Criteria:	Possible Points	Self-Assessment	Teacher Assessment
	10		
	10		

Rate each category according to the following scale:

9-10 = excellent, 7-8 = very good, 5-6 = good, 3-4 = satisfactory, 1-2 = poor, and 0 = unsatisfactory.

Rubric Template #8

(Student Self-Evaluation)

Student's Name: _____				
Class	Activity:	Date:		
_____	_____	_____		
EVALUATION STATEMENTS		RATING		
		1	2	3
		Excellent	Good	Fair
1.				
2.				

Rubric Template #9

Content – Writing (40 points)

Points	Description

Content – Technical (33 points)

Points	Description

Communication (15 points)

Points	Description

Technical Organization (12 points)

Points	Description

TOTAL POINTS = 100

Rubric Template #10**Student's Name:** _____ **Date:** _____**Checklist for** _____**Yes** **No**

_____	_____	_____
_____	_____	_____
_____	_____	_____

Rubric Template #11**Scoring Guide for** _____**Student's Name:** _____ **Date:** _____

Standard: You must achieve a rating of 2 for each of the grading criteria in order to pass the competency. If you fail to do this you will have the opportunity to do extra assignments to work on the area that is lacking.

Criteria	Values
	3 2 1
	3 2 1
	3 2 1
Overall Rating	

Comments:**Evaluator:** _____ **Date:** _____**Key:**

3 = You have met and exceeded the criterion.

2 = You have met the criterion.

1 = You have not met the criterion completely

Rubric Template #12

Program Competency	55%	65%	75%	85%	95%

CTSO EVENTS: A Possible Assessment Vehicle

Simulations and individual competitions are another venue for performance demonstration and assessment. The CTSO Events/Competitions provide a student with the opportunity to demonstrate a specific skill or set of skills under the observation of an external assessor(s), rather than the classroom instructor. Because they are in a competitive situation, students demonstrate skills in order to obtain personal/school recognition for exceptional performance.

Winners in state individual and team competitions participate in national events. Most national events mirror state events, although there are some exceptions particularly in Business Administrative Services, Agriculture, and Family and Consumer Sciences. These program areas have several CTSO events at the national level that are not available in state competitions.

Event winners are determined several ways, ranging from accumulated point values for team events to individual scores in each single event accumulated for higher-level awards (i.e. Outstanding XXX). Some events have extensive standards, criteria and scoring guidelines (rubrics), while others differ significantly.

Some events use only external evaluators; some use a combination of educator and industry representation. Team and chapter events utilize composite member scores rather individual performance scores. For this reason, team events are not considered viable for assessment/certification of competency attainment for individual CTE students unless scoring procedures were changed.

PRO

1. Regularly scheduled events are available for all student participants who chose to be members of the CTSO and are certified as eligible for the competitive event by their instructor.
2. Many events use pre-defined skill areas and rating sheets/standards against which performance is assessed.
3. Simulates a real work environment in some, but not all, events.
4. For participating students, mastery in simulated, situational events is documented by participant scores and recognized by chapter advisors.
5. Documentation of performance exists through rating sheets provided by each assessor/rater for each individual involved in non-team events.

CON

1. Inter-rater reliability is inconsistent (between and among judges).

2. Many events lack scoring rubrics, leading to judgmental rather than objective scoring practices between/among judges.
3. Not all students have the opportunity to attend competitions (lack of available funding and/or are not a member of the respective CTSO).
4. Among the five CTSOs, there is a lack of consistent events for career competencies, particularly cross-program/workplace skills competencies.
5. Many rating scales/scoring sheets are not referenced to the particular state approved competency.
6. Team and chapter events do not rate individual performance. Thus, these events are unlikely to ascertain individual attainment of particular competencies.
7. Several CTSOs revise tests annually (i.e. DECA and FBLA). Thus, each year each new test would have to be matched to Arizona competencies to re-certify the event for competency attainment documentation purposes.

Appendix I contains a matrix of CTSO contests and Cross-Program Competencies. It represents a “first pass” at identifying which events are potentially useful for competency attainment validation. The matrix is lengthy and has had only a preliminary review by the respective CTSO advisors. No attempt has been made to review the event guidelines on an item-for-item basis or to verify that the cross-program competency is, in fact, included in that event and that there are appropriate rubrics to assess the competency.

By definition, any student who participates in a CTSO event automatically fulfills at least part of the CTE Cross-Program Competency #9.0 “Participate in leadership activities such as those supported by career and technical student organizations.” For this reason, the matrix does not address this Cross-Program Competency.

Individual CTSO event content, judges selection practices, event administration guidelines, and scoring rubrics should be cross-referenced to the respective instructional program(s), cross-program competencies, and/or work place skills to determine applicability for vocational competency assessment purposes. At face value, many do not appear to adequately judge competency attainment for accountability and performance measure purposes.

PROFESSIONAL DEVELOPMENT

Vernon Law is credited with the maxim “Experience is the worst teacher: it gives the test before presenting the lesson.”

Behuniak¹¹¹ has proposed three areas for professional development related to tests and test results:

¹¹¹ Behuniak, Peter. (2002). Phi Delta Kappan *Consumer-Referenced Testing*. PDK 84/3. Bloomington, IN. 203.

1. Understanding of the salient features of different tests and basic measurement principles.
2. Familiarity with the specific attributes and purposes of assessment programs directly affecting one's students.
3. Facility in making intelligent use of available test formats/types.

Clearly, part of the process involves changing pre-service teacher education programs to improve understanding test development and test administration practices. However, if the Department or a district wishes to pursue developing and administering more effective criterion-referenced tests and other alternative assessments there is a need to improve teacher's test development and utilization skills.

Norris and Croft¹¹² reiterated seven principles from Watts and Castle (1993) for effective professional development experiences. They include:

1. ...are driven by a well-defined image of teaching and learning.
2. ...provide opportunities for teachers to build their knowledge and skills.
3. ...use or model the strategies teachers will use with their students.
4. ...build a learning community.
5. ...support teachers to serve in leadership roles.
6. ...create links to other parts of the educational system.
7. ...are continually assessed and improved.

Characteristics of effective professional development include that it is ongoing and collaborative. It focuses on student learning, is rooted in the knowledge base for teaching, provides adequate time and follow-up support, and is accessible and inclusive. Professional development activities to encourage academic/vocational curriculum integration, improve teaching strategies and assessment practices, and to strengthen professional relationships are critical components in any change model. If the Department wishes to change curriculum and assessment practices, it should initiate a series of workshops to include:

- Selecting and using existing tests effectively.
- Selecting test types and designing appropriate test items.
- Developing test writing skills.
- Analyzing and validating previously administered test items.
- Using test results to focus instruction.
- Aligning test content and curriculum.
- Testing integrated content knowledge.

To provide professional development technical skills development, the Department should:

1. Set aside fiscal resources to support assessment skills development workshops for a three-year period minimally.

¹¹² Norris, Carol A., and Croft, Vaughn (2001) *Curriculum Design Process and Materials Format*, Arizona Department of Education, Phoenix, AZ. 86

1. Set aside fiscal resources to support on-going ADE and local district technical skills/test item construction and validation workshops for a three year period minimally.
2. Set aside fiscal resources to support continuous review/updating of Curriculum Framework assessment components for a three-year period minimally.
3. Provide joint academic/vocational educator workshops to develop academic/vocational integrated instructional units and distribute these units for statewide use.
4. Identify appropriate test types and construct items to assess Cross-Program Competency attainment (or)
5. Select/develop a single-purpose test to assess Cross-program Competency attainment and establish procedures to assure
 - a) common practices when administering the test,
 - b) funds to support per student costs, and
 - c) procedures and database support to maintain attainment records.

ALTERNATIVE ASSESSMENT IMPLEMENTATION STRATEGIES

Assessment materials developed by educational institutions, state departments of education, consortiums, and industry sources identify parameters for alternative assessment materials development. A composite list of parameters has been developed by the Wisconsin Department of Public Instruction "Business Cooperative Skill Standards Curriculum" (1998) and by LJ Technical Systems "IT2000 Information Technology Program" (2001). Many vendors also have materials such as those available from Microsoft Office "User Specialist Courseware" (1999). Oklahoma, New York, and Ohio Departments of Education and NOCTI and V-TECS have also developed excellent resource guides, assessment materials, and samples.

1. Minimum parameters for content in alternative assessment materials and practices include:

A. Competency Indicators

1. Assess the specific competency/indicator to be attained.
2. Include problem-based, authentic performance tasks, whenever possible.
3. Utilize higher-order and inquiry-oriented activities, whenever possible.
4. Identify the difficulty level of the expected performance (i.e. scale 1-5).
5. Indicate the curriculum level to be attained (i.e. 1=introduction, 2=expansion, 3=specialization).
6. Provide ability criteria and scoring values to be utilized.
7. Assure tests are fair, reliable, valid, and appropriate measures.

B. Administrative Practices

1. Establish learner directions for the assessment process.
2. State rubrics to be followed.
3. Provide a scoring guide for evaluator use (i.e. outstanding, highly successful, not yet successful).
4. Assure that all scoring sheets are signed and dated by both the student and the evaluator.
5. Utilize computer technology for administration and record keeping, as available.
6. Provide opportunities for student feedback, reflection, and redirection.
7. Develop a calendar for periodic revision of teacher/district test materials.

2. Local District Prepared Assessment Materials: Districts wishing to develop alternative assessment materials in lieu of an industry/vendor or state-sponsored assessment for district and state accountability reporting purposes should formalize a development process. Suggested approaches include:

- A. Assemble a multi-teacher development team to draft assessment test items, rubrics, scoring guides, and/or checklists with criteria to use to certify competency attainment. Districts/programs may utilize a variety of assessment types such as:
- ❑ Standardized tests administered by a testing service.
 - ❑ Performance observations/checklists with scoring criteria and values.
 - ❑ Portfolios with teacher developed progress notes on student performance and student self-assessments.
 - ❑ Teacher developed tests (any test type) and scoring guide(s)/rubrics.
 - ❑ Skills checklists with scoring criteria and values.
 - ❑ CTSO competition materials and scoring criteria/rubrics, and values.
 - ❑ Apprenticeship program standards and performance indicators.¹¹³
- B. Analyze item content/structure to assure validity, reliability, and usability.
- C. Certify that the proposed assessment materials and strategies are appropriate to measure 80% or more of the competencies contained in the respective curriculum(s).
- D. Provide a timeline for District review and periodic modification of the assessment(s).
- E. Submit copies of alternative assessments to ADE for certification.

It is important to note that the Arizona Department of Education has not yet determined if/how alternative and in lieu of state recommended assessments would be incorporated into the State accountability and student performance reporting systems. If/when ADE determines they will accept alternative assessments, any district/school wishing to establish these would need to have approval by ADE prior to implementation.

¹¹³ Available through the Arizona State Apprenticeship Council offices. Apprenticeship programs require a high school diploma or GED and the applicant must be 18 years of age prior to application for entrance into the apprenticeship. Programs are minimally 8,000-10,000 hours in length, full time, and with a 3-6 month probationary period. Students participating in a Tech Prep 2+2 program should be made aware of apprenticeship opportunities that may align with their curriculum specialization area.

Summary

This report has addressed a variety of public/professional testing issues, test types and uses, performance measures and accountability reporting, district assessment and data management capabilities, and student assessment practices. Essentially, there are six questions to resolve as the State moves toward improving its accountability system, districts move closer to meeting state performance measures, and student assessment practices are revised/improved. These questions include:

How do student performance assessment practices continue to improve the learning environment and simultaneously accomplish accountability and state performance measures reporting?

What steps can be accomplished to resolve local district/state issues that surround accountability versus assessment practices?

How do districts implement and absorb costs associated with using mandated industry/vendor standards and certifications for CTE student assessment and State reporting purposes?

What are the appropriate uses of various test types and are they acceptable as “in lieu of” tests for state accountability and performance measures reporting?

How can Arizona implement an assessment and performance reporting system that is acceptable to both local districts and the State Department of Education, Career Technical Education Division?

Should a common assessment for Work Place Skills and Cross-Program Competencies be developed or should the State develop test items to include these skills and competencies in the AIMS test?

Conclusions

- Organizations can be divided and grouped into similar categories: Assessment/Test providers, Curriculum developers (with some assessment capabilities potentially), and Standards developers.
- There is limited commonality in materials prepared/available from consortiums, states, organizations, and the U. S. Department of Education with regard to performance standards, academic/vocational curriculum integration models, and assessment strategies.
- The SCANS competency areas are still widely regarded as a standard for assessing workplace readiness skills, although not all states and consortiums have adopted the SCANS skill areas.
- Some States (Oregon, New York, Arkansas, Ohio, Oklahoma, Wisconsin) have made significant strides to integrate academic and technical skill standards into common

curriculum goals with career-related learning standards. Use of alternative assessments are a component of these initiatives.

- Common (core/cross-program) competencies are beginning to be articulated by varying States and consortiums, although most can't/won't articulate them specifically, unless one's state is a member of the pilot group or consortium.
- Career Clusters are also regarded as another set of standards around which curriculum and assessments are organized. Significant data is available from <http://www.careerclusters.org>.
- According to the Education Commission of the States¹¹⁴ Maryland is the only state that has established World of Work and Survival Skills for inclusion in the state's core assessment system.
- Two of the "cross-program competencies" adopted by Arizona are not reflected in the SCANS and/or CTE Cross-Program Competencies (namely Evaluate the role of small business in the economy and Business/financial management for entrepreneurs), nor are they typically assessed for in the myriad of assessment options available through testing sources and consortiums.
- Two of the Arizona Cross Program Competencies (Develop an individual career plan and Evaluate the role of small business in the economy) do not cross-reference to either SCANS or the Arizona Work Place Skills.
- "Develop an individual career plan" is part of the process to "Prepare for employment" and could be combined with it, rather than continued as a separate Cross-Program Competency for Arizona programs.
- Based on limited information, it appears that "Evaluate the role of small business in the economy" as a separate Cross-Program Competency should be reconsidered.
- Review of the crosswalk clearly demonstrates that the Academic Standards essentially already include the CTE Cross-Program Competencies. Assuming the crosswalk satisfies criteria of "adequacy and accuracy," there is little, if any, need to develop additional competencies/indicators
- There is no standard practice regarding the use of skill certificates and industry credentials for secondary and community college vocational/technical education students.
- Districts are generally unwilling to replace their existing assessments with a State model and are more willing to select an alternative assessment model acceptable to both them and the State in order to maintain "approved program" status for accountability purposes in Arizona.

¹¹⁴ ECS Clearinghouse Notes, Advanced Placement Courses and Examinations. (January 2000). Education Commission of the States, Denver, CO.

- There are wide differences among the states in the degree to which Cross-Program Competencies, Workplace Skills, and Technical Standards are included/excluded from a particular state's core assessment and reporting system(s).
- There is little commonality among Arizona school districts with respect to performance assessment systems, test type utilization, administrative guidelines, and testing practices.
- Database capabilities to track and report student performance and to respond to ADE performance indicators differs significantly among Arizona districts ranging from hand posting to sophisticated electronic data management.
- Some districts purchase tests/test services, but per student costs have prohibited many districts from purchasing such industry/vendor-prepared assessments.
- The options of developing an Arizona test, adopting another state's test(s) or industry-endorsed tests, and/or purchasing vendor testing services has not been resolved for CTE technical skills assessment.
- The Arizona AIMS test does not include Workplace Skills, although they are included in the Arizona Standards.
- The Arizona Workplace Skills, SCANS Skills, and all but two CTE Cross-Program Competencies are closely aligned and they, in turn, are adequately included in the Arizona Academic Standards
- Integrated academic and vocational/technical curriculum practices in Arizona have the potential to improve academic and workplace skills development.
- No single test type emerges as "most preferred" for vocational/technical assessments, although performance-based assessment is widely used in Arizona and other states.
- District-level professional resources to assist teachers in developing test items and test administration skills are available in limited manner.
- Test development and test item writing skills require higher levels of technical competency than may exist in some school districts.
- Many districts retain on-site staff and expend large amounts of personnel, time, and money to keep reporting requirements and instructional content and assessment practices current through professional development activities.
- Because of group rather than individual scoring practices, team and chapter CTSO events should not be considered for individual student vocational competency assessment and accountability reporting.

- It is questionable that individual CTSO event content, judges' selection practices, event administration guidelines, and scoring rubrics are cross-referenced to the respective instructional program(s), cross-program competencies, and/or work place skills to determine vocational competency assessment.
- Recently implemented Design Team makeup, procedures, and content requirements in the ADE Curriculum Frameworks provide considerably more industry/vendor assessment and certification resource information for Arizona CTE teachers than in past curriculum guides.
- New Curriculum Frameworks follow, for the most part, common formats; however, uncommon formats have been utilized when referencing available assessment and certification sources.

Recommendations

1. Conduct an analysis of the completed curriculum Design Team crosswalks to check for consistency with those contained in this report's Cross-Program Competencies and Academic Standards to modify, if necessary, the Framework recommendations.
2. Combine "Develop an individual career plan" with "Prepare for employment" rather than continuing use of it as a separate Cross Program Competency.
3. Reconsider and eliminate or reconfirm the appropriateness of "Evaluate the role of small business in the economy" and "Business/financial management for entrepreneurs" as Cross Program Competencies.
4. Crosscheck the various Design Team assessment lists to eliminate duplicates within the recommended certification(s) so as to identify/prioritize "preferred" industry certifications to adapt/adopt for use in each program.
5. Review differences in program option lists and consider adopting existing assessments, developing a new assessment, and/or reaching compromise between CTE programs to select a common Cross-Program Competencies assessment test.
6. Decide whether to develop a state-sponsored Arizona Cross-Program Competencies test, adopt/adapt another district or state's test, or purchase testing services from a vendor. **(or)**
7. Develop and field-test recommended Workplace/Cross-Program Competencies test items to incorporate into the AIMS standardized test prior to implementation of testing cycles.
8. Develop a strategy to assure and validate that career/technical students acquire Work Place Skills and Cross-Program Competencies prior to program completion.

9. Bring people together across school districts and industries to establish testing content for *occupations that overlap* and to develop guidelines to assure the assessment system for these standards is fair, nondiscriminatory, and non-repetitive.
10. Empower and assist teachers, vocational directors/administrators, and CTE staff in developing teaching/learning objectives, instructional materials, and assessment instruments that meet State “acceptability” criteria.
11. Review and adopt recommended implementation guidelines to utilize locally developed assessments as “in lieu of” tests subject to ADE approval.
12. Examine NOCTI tests for potential statewide use in CTE Cross-Program Competencies and program technical skills assessment.
13. Explore ADE capability to absorb per student costs statewide for vendor-prepared assessments and industry-endorsed credentialing exams.
14. Establish a study team to address strategies to implement common reporting practices among the states’ school districts, including a reporting mechanism to use for federal and state accountability purposes.
15. Align district data collection efforts with state enrollment and student assessment data reporting requirements.
16. Maintain stable, consistent database requirements for local districts to reference ADE accountability and performance assessment reporting for at least three years.
17. Investigate establishing an ADE database for information about “acceptable” industry certifications in CTE program(s) and, if necessary, add a new support staff position within ADE to maintain the database.
18. Resolve issues regarding teacher retraining priorities, appropriateness of various assessment types (objective tests, portfolios, essay, performance, etc.), test procedures (particularly for out-of-level testing and special needs student test accommodation) and intermittent/end-of term student competency assessments that consider varied student learning and performance styles with a work group from the AST.
19. Fund and implement professional development programs to improve instructional materials, practices, and assessments to document that students acquire CTE Cross-Program Competencies and the Arizona Workplace Skills upon program completion.
20. Conduct capacity-building workshops for teachers and CTE professional staff to develop test item writing skills.
21. Negotiate administrative policies with local districts that will support on an on-going basis release time, resources, technical assistance, and professional development

opportunities for local staff to improve and maintain assessment activities and database reporting systems.

22. Provide the academic, cross-program, and workplace skills standards Crosswalk to each curriculum development Design Team with instructions to utilize the index in developing program-specific curriculum/crosswalks, instructional materials, and recommended assessments to address these standards.
23. Develop a standard format for curriculum Design Teams to utilize in presenting industry certification, test sources, and alternative assessment information and require each project in the future to use that format.
24. Exclude team and chapter CTSO events for consideration for vocational competency assessment purposes.
25. Conduct an analysis (cross-reference) of individual CTSO event content, judges selection practices, event administration guidelines, and scoring rubrics to assure they adequately assess competency attainment for the respective program competencies.
26. Consider an assessment guideline like that of Arkansas whose guideline states 1) look for industry certification, 2) if not available, then use NOCTI 3) if not appropriate, then utilize V-TECS test item bank, and 4) if not available or appropriate, then develop district/school-based assessment materials.
27. Adapt/adopt the PACT-Alt Scoring Rubric for special populations as a rubric model for Arizona teachers to utilize.
28. Implement use of teleconferencing and/or distance learning to conduct information seminars to inform local districts of procedural changes prior to implementation by ADE.
29. Join at least two consortia that specialize in vocational standards, curriculum, instructional materials, and assessment activities and adopt/adapt their materials for Arizona CTE programs, as applicable.
30. Establish guidelines to address test administration practices and content issues of reliability (assessment instruments that are administered consistently according to national or state reliability standards), scope ((statewide policies to ensure that attainment is measured appropriately in all schools) and alignment (assessments aligned to state-established industry-validated standards).

References

- A SCANS Report for America 2000. Department of Labor, Washington, DC: The Secretary's Commission on Achieving Necessary Skills.
- American College Testing Service: *ACT Policy for Documentation to Support Requests for Testing Accommodations on the ACT Assessment* website:
<http://www.act.org/aap/disab/policy.html>
- Ananda, S. M. et al. (1995) "Skills for Tomorrow's Workforce." *Policy Briefs*, no. 22. San Francisco: Far West Lab for Educational Research and Development, December (ED 392 132)
- Arizona CareerInfoNet @ http://www.acinet.org/acinet/lois_agency.htm?stfips=04&by-state&x=28&y=9
- Arizona Performance Measures. Secondary FY2002 Guidelines for Career and Technical Education Program Evaluation. (Revised January 2002)
- Asche, M. (1990) "*Standards and Measures of Performance: Indicators of Quality for Virginia Vocational Education Programs.*" Paper prepared for the teleconference "Preparing a Competent Work Force through Indicators of Quality for Vocational Education." Blacksburg: Division of Vocational and Technical Education, Virginia Polytechnic Institute and State University 6.
- Assessing Achievement in Home Economics Education*, New York State Education Department, Albany, N.Y. 1991.
- Bailey, T. and Merritt, D. *Making Sense of Industry-Based Skill Standards*.
- Berkeley, CA: National Center for Research in Vocational Education (1995) (ED 389 897)
- Balogh, Judy; Crary, Michelle and Libette Garcia, Chris. *Rubric Maness! Student and Program Assessment Made Easy For Business Education*" Paper presented at the Arizona State Annual Vocational Conference, July 22, 2002. Tucson, Arizona. Copies available from Ms. Judy Balogh (Arizona State University) and/or Dr. Janet Gandy (Arizona Department of Education, Career and Technical Education Division).
- Behuniak, Peter. (2002). Phi Delta Kappan *Consumer-Referenced Testing*. PDK 84/3. Bloomington, IN. 201.
- Boyer, Ernest. <http://www.ed.gov/offices/OVAE/CTE/2pgperk.html>
- Bradley, Ann. (2000, July 12). Union heads standards warnings. *Education Week*, 7(12), 1,20-21.
- Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27. No. 1) Washington, DC: The George Washington University, Graduate School of Education and Human Development.

Chicago Public Schools Performa Online Resource for RUBRICS (Rubric Bank) @ [http://intranet.cps.k12.il.us/Assessments/Ideas and Rubrics/Rubric Bank/rubric_bank.html](http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Rubric_Bank/rubric_bank.html)
A complete copy of their web site resource list is included in Appendix XXX. A link is also available through <http://www.napehq.org/>

Clagett, C. A. (1997) *Workforce Skills Needed by Today's Employers. Market Analysis MA98-5*. Largo, MD: Prince George's Community College, Office of Institutional Research and Analysis (ED 413 949) op.cit. and Oliver, K. M.; Russell, C.;

Custer, Rodney L., Schell, John, McAlister, Brian D., Scott, John L., and Hoepfl, Marie "Using Authentic Assessment in Vocational Education" ERIC/ACVE IN 381. 2

Gilli, L. M.; Hughes, R. A.; Schuder, T.; Brown, J. L.; and Towers, W. (1997) "Skills for Workplace Success in Maryland: Beyond Workplace Readiness." In *Workforce Readiness: Competencies and Assessment*, edited by H. F. O'Neil, Jr. Mahwah, NJ: Lawrence Erlbaum.

Drummond, I.; Nixon, I. and Wiltshire, J. "Personal Transferable Skills in Higher Education: The Problems of Implementing Good Practice." ***Quality Assurance in Education*** " no. 1 (1998):19-27, 21

Ebel, Robert L Measuring educational achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965, Chapters 4-6, and Ebel, Robert L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972, Chapters 5-8.

ECS Clearinghouse Notes, Advanced Placement Courses and Examinations. (January 2000). Education Commission of the States, Denver, CO

Elmore, R. F., Abelman, C. H., & Fuhrman, S. H. (1996) The new accountability in state education reform: From process to performance. In H. F. Ladd (ed.) *Holding Schools accountable: Performance-based reform in education* (pp.93-94). Washington, DC: The Brookings Institution.

Employability Skills Certificate Program, Lifework Education Team, Department of Public Instruction, P. O. Box 7841, Madison, WI 53707-7841, Fax: 608-267-9275

ERIC Clearinghouse on Assessment and Evaluation (ERIC/AE) links to several web sites, including *Scoring Rubrics – Definitions & Constructions* available on line @ http://ericae.net/faqs/rubrics/scoring_rubrics.htm

Frary, Robert B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation*. 4(11); <http://ericae.net/pare/getvn.asp?v=4&n=11>. or ERIC ED 398238 ERIC Clearinghouse on Assessment and Evaluation, Washington, D.C.

Geber, B. "The Plan to Certify America." Training 32, no. 2 (February 1995): 39-42, 44.

Gronlund, N. E. (1982) Constructing achievement tests. 3rd ed. Englewood Cliffs, JM: Prentice-Hall.

Guile, D. "Skill and Work Experience in the European Knowledge Economy." *Journal of Education and Work* 15, no. 3 (September 2002): 268-269.

Haladyna. T. M. (1999) Developing and validating multiple-choice test items. 2nd ed. Mahwah, NJ: Erlbaum.

<http://ericae.net/pare/getvn.asp>? (web site for the ERIC Clearinghouse on Assessment and Evaluation), <http://www.oir.uiuc.edu/dme/exame/ITQ.html> (web site for the University of Illinois Urbana-Champaign and at <http://www.use.umn.edu/oms/multchoice.htmlx> (web site for the University of Minnesota, Office of Measurement Services).

Imel, Susan. (1999) *Work Force Education: Beyond Technical Skills* Trends and Issues Alert No. 1ERIC/ACVE

Kehoe, Jerard (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation*. 4(9) <http://ericae.net/pare/getvn.asp?v=4&n=9>; or ERIC ED398236 http://www.ed.gov/databases/ERIC_Digests/ed398236.html;

La Marca, Paul M. (2001) *Alignment of Standards and Assessments as an Accountability Criterion* ERIC Clearinghouse on Assessment and Evaluation, ISSN 1531-7714. 6

Lankard Brown, Bettina. (2002) *Generic Skills in Career and Technical Education Myths and Realities* No. 22, ERIC/ACVE. Washington, DC

Lankard Brown, Bettina. (1997) *Skill Standards: Job Analysis Profiles Are Just the Beginning* Trends and Issues Alert ERIC/ACVE

Lankard Brown, Bettina.(2002) *Skill Standards: Job Analysis Profiles Are Just The Beginning* Trends and Issues Alert. ERIC/ACVE, 1997

Losh, Charles L. (2002) *Using Skill Standards for Vocational-Technical Education Curriculum Development Information Series* No. 383 ERIC/ACVE, Washington DC

Losh, Charles L. (2000) *The Linkage System: Linking Academic Content Standards and Occupational Skill Standards* (Ver 1.2) V-TECS, Southern Association of Colleges and Schools. <http://www.v-tecs.org>

Lynch, Richard L., *New Directions for High School Career and Technical Education in the 21st Century*, ERIC Clearinghouse on Adult, Career and Vocational Education, Center for Education and Employment, Ohio State University, Colombus, OH

McMillan, James H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical assessment, Research & Evaluation*, 7(8). <http://ericae.net/pare/getvn.asp>?

Marzano, Pickering, and McTighe in Scott, John L., Using Authentic Assessment in Vocational Education ERIC/ACVE IN 381. 10

Meier, Deborah. (November 2002) *Standardization vs. Standards*. Phi Delta Kappan 84(3).
Bloomington, IN 192.

Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). Available online:
<http://ericae.net/pare/getvn.asp?v=7&n=25>

Moskal, Barbara M., (2000) *Scoring Rubrics: What, When and How?* Practical Assessment, Research & Evaluation, 7(3) ERIC Clearinghouse on Assessment and Evaluation. Available online: <http://ericae.net/pare/getvn.asp?v=7&n=3>

Murnane, R. J., and Levy, F. (1996) *Teaching the New Basic Skills. Principles for Educating Children to Thrive in a Changing Economy*. New York: Free Press

NOCTI, *Using Standardized Test Data To Improve Instruction In Career-Technical Education, A Perspective for Practitioners*. (undated), 4.

Norris, Carol and Croft, Vaughn. (2001) *Curriculum Design Process and Materials Format*, Arizona Department of Education, Phoenix, AZ.

North Central Association, <http://www.ncacsi.org/transitions/>

Olson, Lynn. (2000a). High-stakes tests jeopardizing Hispanics, panel warns. *Education Week*, 7(12), 7.

Olson, Lynn. (2000b) Test-makers' poll finds parents value testing. *Education Week*, 8(2), 16.

Parsad, Basmat; Farris, Elizabeth. "Occupational Programs and the Use of Skill Competencies at the Secondary and Postsecondary Levels," 1999, NCES 2000-023, Bernie Greene, project officer. Washington, DC 1.

Penn State University Testing Services (2000) Academic Testing Test Design and Construction. http://www.uts.psu.edu/Test_construction_frame.htm

Pucel, D. J. "The Changing Roles of Vocational and Academic Education in Future High Schools." Paper presented at the Central Educational Science Research Institute, Beijing, China, October 4, 1999. (ED 434 242)

Rahn, M.L; O'Driscoll, P.; and Hudecki, P. (1999) *Taking off! Sharing State-Level Accountability Strategies: Using Academic and Vocational Accountability Strategies to Improve Student Achievement*. Berkeley, CA: National Center for Research in Vocational Education (ED 431 138)

Richens, G. P., and McClain, C. R. "Workplace Basic Skills for the New Millennium." *Journal of Adult Education* 28, no. 1 (Summer 2000): 29-34

Robertson, Anne S. "High-Stakes" Testing: New Guidelines Help Direct School Change. *NPIN Parent News* for November-December 2000. 2
<http://npin.org/pnews/2000/pnew100/intl100b.html>

Sandy Pritz, NOCTI, 11/27/02 email

Scott, John (2002) *Authentic Assessment Tools* The University of Georgia. ERIC/ACVE, IN 381.2. 33-48

Sevenair, John P., Item Writing guidelines, Xavier University,
<http://webusers.xula.edu/jsevenai/objective/guidelines.html>

Shepard, L. A. (2000). The Role of Assessment in a Learning Culture. Paper presented at the Annual Meeting of the American Educational Research Association. Available
<http://www.aera.net/meeting/am2000/wrap/praddr01.html>

South Carolina Palmetto Achievement Challenge Tests Alternate Assessment Portfolio Guide, South Carolina State Department of Education. <http://www.sde.state.sc.us>

Stecher, et al (1997) "*The Cost of Performance Assessment in Science: The Rand Perspective*" Paper presented at the Annual Meeting of the National Council on Measurement in Education San Francisco, CA (April 1995) ERIC No. ED 383 732.

Stevens, David W. (2001). *21st Century Accountability: Perkins III and WIA Information Paper 1002*. National Dissemination Center for Career Technical Education. Ohio State University. Columbus OH.

The National Forum on Assessment: Principles and Indicators for Student Assessment Systems. (1995) National Center for Fair and Open Testing (Fair Test), Cambridge, MA

University of Illinois Urbana-Champaign "Improving Your Test Questions" Urbana, IL.
<http://www.oir.uiuc.edu/dme/exams/ITQ.html>

University of Minnesota, Office of Measurement Services (1999) "Writing Multiple-Choice Items" <http://www.ucs.umn.edu/oms/multchoice.htmlx>

US Department of Education, National Center for Education Statistics (2000) *E.D.Tab*,
US Department of Education, Office of Vocational and Adult Education, (2001)
<http://www.ed.gov/ovae>

Werner, M. C. (1995) *Australian Key Competencies in an International Perspective*.
Leabrook, Australia: National Centre for Vocational Education Research, (ED 407 587)

Whetzel, Deborah, (1992) "The Secretary of Labor's Commission on Achieving Necessary Skills," Eric Digests ED339749. (http://www.ed.gov/databases/ERIC_Digests/ed339749.html)

Willis, Scott. Education Update, (1999) "*The Accountability Question*," Association for Supervision and Curriculum Development, Vol. 41, Number 7, 1.

Willis, Scott. (1999) *The Accountability Question* Education Update ASCD 41(7) Alexandria, VA. November 5

Wills, J. (1997) *Standards: Making Them Useful and Workable for the Education Enterprise*. Washington, DC: Office of Vocational and Adult Education. US Department of Education (ED 431 461)

Wills, Joan. (2002) *Promoting New Seals of Endorsements in Career Technical Education* The National Association of State Directors of Career Technical Education Consortium, Washington, D.C. 8

Wonacott, Michael E., *Standards: An Embarrassment of Riches* In Brief: Fast Facts for Policy and Practice National Dissemination Center for Career & Technical Education Washington: DC 2000 (4) I

Penn State University Testing Services, (2000), Academic Testing Test Design and Construction. University Testing Services.
http://www.uts.psu.edu/Test_construction_frame.htm

<http://ericae.net/pare/getvn.asp>? ERIC Clearinghouse on Assessment and Evaluation) and at <http://www.use.umn.edu/oms/multchoice.htmlx> (University of Minnesota, Office of Measurement Services)

Penn State University Testing Services (2000) Academic Testing Test Design and Construction. http://www.uts.psu.edu/Test_construction_frame.ht

WEB RESOURCES

<http://www.acinet.org> - Wages and employment trends, occupational requirements, state by state labor market conditions, and extensive online career resource library.

<http://www.act.org/workkeys/education/works.html> WorkKeys website

http://www.aztechprep.org/Vocational_Programs/vocational_programs.html -AZ site with rubrics, standards

<http://www.ed.gov/offices/OVAE/HS/gray.doc> - US Office of Vocational and Adult Education

<http://www.emsc.nysed.gov/workforce/cte/nationalassess.htm> - NY State Education Dept.

<http://www.ERIC.ACVE.com> - General research and resources site

<http://www.ericae.net/edo/ED318915.html> - An excellent summary of sources for curricula, materials, assessment and testing information.

<http://www.ihdi.uky.edu/mcrrc/> - Mid-South Regional Resource Center

<http://www.myschools.com> - Site with links to most state standards sites

<http://www.nccte.org> - National Center for Career and Technical Education

<http://www.nchste.com> - Test items for industry validated standards.

<http://www.ode.state.oh.us/ctae/principal/assess/default.asp> - Ohio Department of Education

<http://www.pen.k12.va.us/VDOE> - Virginia Department of Education

<http://www.rcmp.org> - Canadian site with extensive core and specialized competencies for law enforcement

<http://www.usoe.k12.ut.us> - Utah Office of Education

<http://www.vtecs.org> - Resource site on clusters, competencies, etc.

RUBRIC RESOURCES

Build your own rubrics online:

http://landmark-project.com/classweb/tools/rubric_builder.php3

http://www.teach-nology.com/web_tools/rubrics/general/

<http://rubistar.4teachers.org/index.shtml>

<http://www/rubricbuilder.on.ca/learn.shtml>

Assessment and Rubric Sites:

http://www.uni.edu/profdev/teachnet/four/eval_g4.html

BlueWeb 'n Rubric:

<http://www.kn.pacbell.com/wired/bluewebn/rubric/html>

Chicago Public Schools Performance Assessment Ideas:

http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/ideas_and_rubrics.html

Collaboration Rubric:

<http://edweb.sdsu.edu/triton/tidepoolunit/Rubrics/collrubric.html>

Categorized, annotated list of over 2600 sites to enhance instruction and support the curriculum:

<http://school.discovery.com/schrockguide/assess.html>

Performance assessments for history, science, and mathematics, and the CRESST Scoring Rubric:

<http://www.cse.ucla.edu/CRESST/pages/samples.html>

Portfolio Web Page Rubric:

<http://www.cho.arizona.edu/inst/edp512297/portfoliowp.html>

<http://www.sdcoe.k12.ca.us/notes/5portfolio.html>

<http://www.ed.gov.pubs/OR/ConsumerGuides/admuses.html>

<http://www.eduplace.com/rdg/res/literacy/assess6.html>

<http://www.ncrel.org/sdrs/areas/issues/students/learning/lr2post.html>

Rubrics for evaluating web sites:

<http://edtech.sandi.net/rubric/>

Rubrics for Lesson Plans

<http://edweb.sdsu.edu/webquest/rubrics/weblessons.html>

Rubrics for evaluating writing:

<http://7-12educators.about.com/library/weekly/blrubricindex.html>

Multimedia project rubric:

<http://www.ncsu.edu/midlink/rub.senst.html>

<http://www.learningspace.org/instruct/lessons/pst4.html>

Rubric for PowerPoint presentation:

<http://www.schools.lth5k12.il.us/aviston/KBLesson8.html>

Microsoft Office rubric:

<http://www.ga.k12.pa.us/curtech/stucours/offrubr.html>

Web page design rubric:

http://enternet.lth1.k12.il.us/dupage/Session_1.2/lworchester/default.html

<http://www.portledge.org/Laptops/WebRubric.html>

<http://www.westmoor.district28.k12.il.us/tech/Studentrubric.html>

http://www.uni.edu/profdev/technet/four/eval_g4.html

Rubrics Text Reference: <http://www.scarecroweducation.com>

Rubrics: A Handbook for Construction and Use, edited by Germaine L. Taggart, Sandra J. Phifer,

Judy A. Nixon, and Marilyn Wood. 2001. ISBN 1-56676-652-4. For orders and information contact: Scarecrow Press, Inc.

4720 Boston Way, Lanham, Maryland 20706 Phone: 1.800.462.6420 Fax: 717.794.3803

ASSESSMENTS AND TESTING

<http://Act.org>

ACT is an independent, not-for-profit organization that provides more than a hundred assessment, research, information, and program management services in the broad areas of education and workforce development

Work Keys tests skills in problem solving, communication, and teamwork. It also identifies the skill levels needed to do specific jobs

Work Keys shows students their skill levels in eight foundational skills (the skills needed to learn other skills):

- Applied Mathematics
- Applied Technology
- Listening
- Locating Information
- Observation
- Reading for Information
- Teamwork
- Writing

In addition, the Work Keys Targets for Instruction help educators develop curricula and instructional strategies for the Work Keys skill areas. Targets for Instruction are manuals designed to help educators develop curricula and instructional strategies for the Work Keys skill areas.

By using Work Keys information,

- **learners and workers** can document employability skills.
- **employers** can define the skills they are looking for and identify workers who have them.
- **educators** can tailor instructional programs to help learners acquire the skills employers' need.

<http://i-car.com>

The **I-CAR** Education Foundation was created in 1991. The Foundation is a not-for profit organization that is working to attract quality entry-level candidates and assist in preparing them for careers in the collision industry. The Foundation provides the most advanced curriculum, instructor training research and related services to career and technical education.

<http://www1.faa.gov/>

The Federal Aviation Administration

Provides assessment and certification for airline-related careers. Also provides curriculum guidelines.

<http://ets.org>

Educational Testing Service

Educational Testing Service is the world's largest private educational testing and measurement organization and a leader in educational research. ETS has an extensive library of 20,000 tests and other measurement devices from the early 1900s to the present.

<http://nces.ed.gov>

National Center for Education Statistics

NCES collects and reports information on the academic performance of the nation's students as well as the literacy level of the adult population. The National Assessment of Educational Progress (NAEP) is NCES' primary assessment of what American elementary/secondary students know and can do in academic subjects. This NCES program also assesses the proficiency of adults in performing basic literacy and mathematical tasks.

<http://brainbench.com>

Brainbench provides online assessment and certification of over 400 different skills that drive business success today.

<http://www.iteconline.org>

Independent Technicians Education Coalition

ITEC is a volunteer and non-profit entity dedicated to developing and maintaining world class technical training programs for entry level and professional level automotive personnel.

<http://asecert.org>

National Institute for Automotive Service Excellence

Founded in 1972, ASE has a single mission: To improve the quality of automotive service and repair through the voluntary testing and certification of automotive technicians.

<http://asashop.org>

Automotive Service Association

Training and certification of automotive service professionals.

<http://nocti.org>

The National Occupational Competency Testing Institute

NOCTI is a leading provider of occupational competency assessments and services. NOCTI's products and services include job and task analysis, test development, written and performance assessments, scoring services and specialized reporting. Clients have the option of selecting from over 170 standardized technical tests in a variety of occupational fields or customizing assessments for their specific needs.

<http://www.alignmark.com/>

AlignMark

The AccuVision Systems evaluate a candidate's skills and abilities that are required for success in a specific job position. The AccuVision Workforce Readiness System is a unique assessment tool that uses job simulation, video and computer technologies to capture the skills and abilities required for success in customer care and a variety of customer contact, entry level positions. Skills assessed include Customer Relations, Decision Making, Commitment to Quality, Personal Qualities, Responsibility, Self-esteem, Self-management, and Sociability.

STANDARDS

<http://acteonline.org>

Association for Career and Technical Education

The Association for Career and Technical Education is the largest national education association dedicated to the advancement of education that prepares youth and adults for careers. Offer limited products and assessments.

<http://nssb.org>

The National Skill Standards Board (NSSB)

The NSSB is a coalition of leaders from business, labor, employee, education, and community and civil rights organizations created in 1994 to build a voluntary national system of skill standards, assessment and certification systems to enhance the ability of the United States workforce to compete effectively in a global economy. The standards will be based on high performance work and will be portable across industry sectors. The NSSB has categorized the workforce into 15 industry sectors

<http://www.mcrel.org/standards-benchmarks/>

Mid-continent Research for Education and Learning

For over a decade, McREL has been in the forefront of research, practice, and evaluation related to standards-based education. Among the most notable of McREL's contributions to the field is a compendium of K-12 and Career Education standards. In addition, McREL has authored a wide variety of publications and products related to standards, many of which can be downloaded from this site.

<http://nchste.org>

National Consortium on Health Science and Technology Education

Organized in 1991, its purpose is to stimulate creative and innovative leadership for ensuring a well-prepared health care workforce. Primary strategies include fostering collaboration among educational agencies, the health care community, policy-making bodies, and labor.

<http://skillsusa.org>

SkillsUSA-VICA

Skills USA–VICA is a national organization serving more than 250,000 high school and college students and professional members who are enrolled in training programs in technical, skilled, and service occupations, including health occupations.

The *Total Quality Curriculum* enhances Skills USA's Quality at Work movement by preparing students for the world of work starting in the classroom. The curriculum emphasizes the competencies and essential workplace basic skills identified by employers and the U.S. Secretary of Labor's Commission on Achieving Necessary Skills (SCANS).

<http://nasdvtec.org>

National Association of State Directors of Career Technical Education Consortium

The association's mission is to provide leadership for career technical education's role in educational improvement, workforce preparation and economic development.

<http://www.nactei.org/>

National Association for Career Technical Education Information

For over two decades, NACTEI (formerly AVIA) has served as an open forum for the exchange of ideas and methods relating to career-technical education information systems (i.e. data collection, reporting, information management, and fiscal transactions) that are associated with accountability and program improvement.

<http://v-tecs.org>**VTECS - A Consortium for Innovative Career and Workforce Development Resources**

Since 1973, VTECS has operated as a consortium of states where members pool resources to develop research-based information and resources for career and workforce development. Member states include: Alabama, Arizona, Arkansas, Florida, Georgia, Hawaii, Idaho, Illinois, Indiana, Maryland, Minnesota, Missouri, New Hampshire, New Jersey, Rhode Island, South Carolina, Vermont, Virginia, Wyoming. Additionally, the US Air Force, Army, Coast Guard, Marine Corps, and Navy are members.

<http://nccte.com>**National Dissemination Center for Career and Technical Education**

The center has specific projects and activities, which it is focused on annually. Additionally, it provides thought-leadership through its professional speakers web casts.

<http://www.careerclusters.org>**Career Clusters**

Career Clusters provide a way for schools to organize instruction and student experiences around sixteen broad categories that encompass virtually all occupations from entry through professional levels. Resources such as KNOWLEDGE AND SKILLS STRUCTURES and BROCHURES are available for each of the sixteen clusters.

<http://online.onetcenter.org/>**O*Net Online**

O*NET OnLine is an application that was created for the general public-to provide broad access to the O*NET database of occupational information.

The O*NET database includes information on skills, abilities, knowledge, work activities, and interests associated with occupations. This information can be used to facilitate career exploration, vocational counseling, and a variety of human resources functions, such as developing job orders and position descriptions and aligning training with current workplace needs.

Information in O*NET is available for over 950 occupations. Each occupational title and code is based on the most current version (1999) of the Standard Occupational Classification system.

<http://natef.org>

National Automotive Technicians Education Foundation

NATEF was founded in 1983 as an independent, non-profit organization with a single mission: To evaluate technician training programs against standards developed by the automotive industry and recommend qualifying programs for certification (accreditation) by ASE, the National Institute for Automotive Service Excellence.

The NATEF process has resulted in certified automotive training programs in all fifty states at the secondary and post-secondary levels. NATEF also evaluates the providers of in-service technician training programs under a program called Continuing Automotive Service Education (CASE).

CURRICULUM

<http://mavcc.org>

Multi-state Academic and Vocational Curriculum Consortium develops and distributes competency-based instructional materials.

<http://www.rcmp-learning.org>

Royal Canadian Mounted Police Online University

This site is an automated resource data base of learning opportunities including individualized instruction modules, exercises, reading materials; suggested on-the-job assignments, coaching opportunities; and information on formal training including courses and workshops. The site identifies how to locate materials not owned by the RCMP, and provides access to materials designed by the RCMP.

<http://i-car.com>

The **I-CAR** Education Foundation was created in 1991. The Foundation is a not-for profit organization that is working to attract quality entry-level candidates and assist in preparing them for careers in the collision industry. The Foundation provides the most advanced curriculum, instructor training research and related services to career and technical education.

<http://www.cord.org>

Center for Occupational Development

CORD assists educators in secondary schools and colleges through new curricula, teaching strategies, professional development, and partnerships with community leaders, families, and employers. CORD's initiatives include curriculum design, developing new learning tools, delivering professional development, creating applications of educational technology, and conducting educational research and evaluation.

<http://okcareertech.org/cimc>

CIMC – Curriculum and Instructional Materials Center

The Curriculum and Instructional Materials Center (CIMC) is one of the nation's largest developers of competency-based instructional systems. CIMC is a division of the Oklahoma Department of Career and Technology Education.

National Center for Education Statistics**Glossary of Assessment Terms¹¹⁵**

Accountability The demand by legislative bodies, public officials, employers, and taxpayers for school officials to prove the educational impact of the money invested annually in education. This has led to the rise of what is called “accountability testing,” designed to sample what large numbers of students have learned. This is contrasted to “instructional testing” – assessments designed to help teachers improve student learning in the classroom.

Achievement Test A test designed to measure a student’s “school-taught” learning. Usually covers basic skills, such as reading, language arts, and mathematics.

Alternative Assessment Any assessment that is not limited to a pencil-and-paper (norm-referenced or criterion-referenced) or multiple-choice test.

Anecdotal Records A teacher’s collection of observations of a student’s performance; may include dated teacher reflections, checklists, audio/video tapes, photographs, conference, and interviews.

Aptitude Test A test intended to measure a student’s innate ability to learn, given before receiving instruction.

Assess To analyze student accomplishment, usually using a variety of techniques (e.g., performances, teacher observations, teacher performances, teacher observations, scored discussions, portfolios).

Assessment Systematic gathering and synthesizing of information about a person; usually based on various sources of evidence.

Authentic Assessment An alternative assessment method that tests students’ ability to solve problems or perform tasks resembling challenges of the real world. Grant Wiggins writes that authentic assessment must “replicate the challenges and standards of performance that typically face writers, business people, scientists, community leaders, designers, or historians.”

Authentic Task A simulated or “real life” demonstration of learning; examples may include debate, video production, play, experiment, science project, and role-play.

Competency Test A test to determine that a student meets minimum skill and/or knowledge standards, usually for promotion or graduation.

Content Standards Statements that define what students ought to know and be able to do; observable, measurable, or inferable and stated in results-focused terms; reflect broad goals; are comprehensive and developmental.

¹¹⁵US Department of Education, National Center for Education Statistics, “Occupational Programs and the Use of Skill Competencies at the Secondary and Postsecondary Levels, 1999, NCES 2000-023, by Basmat Parsad and Elizabeth Farris. Bernie Greene, project officer. Washington, DC: February 2000, p.1.

SEE ALSO: <http://www.nwline.com> - A link site to Big Dogs Human Resources Development page that has an extensive education glossary in short, easy to grasp definition form.

Criterion-referenced A test in which a student is measured against a given set of criteria. (Compare with the definition for norm-referenced.)

Essay An assignment or assessment that requires students to answer questions emphasizing recall rather than choosing a correct alternative.

Evaluation Judgment of the quality, value, or worth of performance of a program or product.

Exemplars Models that depict desired attributes of quality in ways that students can understand.

Grade Equivalent The grade level at which a student performs on a standardized test. A score of 5.5 means that the student is doing as well as the “average” student in the fifth month of the fifth grade.

Group Processing A cooperative strategy that allows students the opportunity to practice effective listening, responding, and validation skills with others.

High-Stakes Testing Any testing system or activity having important consequences for students, schools, or school districts. Generally high stakes tests are mandated and used for student performance reporting and/or school ratings.

IQ (Intelligence Quotient) Tests Developed more than a century ago, this standardized norm-referenced test supposedly measures a person’s native intelligence. Used more today for psychological screening purposes.

Item Analysis Analyzing each item on a test to determine proportions of students selecting each answer. Can be used to diagnose particular strengths and weaknesses of students, as well as the test’s validity or possible bias.

Kid Watching Formal or informal observation by the teacher of student’s performance and/or interaction. Usually recorded in anecdotal records, checklists, pictures, audio/video tapes.

Learning Logs Students’ reactions to their learning experiences, including insight to process, content, and problems; may include journal entries, written responses to probes, personal reflections.

Multiple-Choice Test Generally a written test in which students are required to complete a statement or answer the question from the alternatives provided.

Norm By definition, norm is the midpoint of performance: 50 percent score above the “norm” and 50 percent below.

Norm-referenced Test A method that relates the score of each student to those in a representative (norm) group; reveals how well each student or group of students did compared to the original group taking the test.

Normal Curve Equivalent (NCE) A normalized standard score with a mean of 50 and a standard deviation of 21.06. The score is most often used to enable the test user to manipulate the test data algebraically.

Outcome General goal statement for student learning.

Percentile Percentile ranks range from a low of 1 to a high of 99 with 50 denoting average performance. The percentile rank corresponding to a given score indicates the percentage of a reference group obtaining scores equal to or less than that score. For example, if a student

scores at the 65th percentile, it means that he or she performed better on the test than 65 percent of the norm group.

Performance Assessment Also called performance-based assessment; an assessment based on pre-established criteria that require a student to perform a task that is observed and judged by raters.

Performance Indicators Statements that specify how knowledge is to be used, or the kind of performance we expect from students relative to desired exit results.

Performance Standards Statements that specify the level or quality of the performance we expect from students relative to desired exit standards. Standards characterize exemplary performance and are set once the task or process and appropriate criteria are established. Performance standards describe how well learners should know or be able to do something.

Portfolio A purposeful collection of student work in a variety of formats providing representative documentation of the learners' efforts, progress, and achievements. Materials demonstrate growth/accomplishment over a period of time and may be centered on a specific topic or content area. Established criteria are used to determine the student's level of performance and criteria for selection. Evidence of student self-reflection is also in the portfolio.

Probe A question or starter statement that promotes personal assessment of the learning process. Probes are often used in learning logs for reflective statements/essays of projects and/or products.

Project A complex assignment that expects more than one type of activity and production for completion. A form of performance assessment.

Quartile After percentiles are determined, the distribution may be broken down for reporting purposes into four groups: the 0-25th percentile, 26-50, etc.

Quintile A similar breakdown, but into five sections: 0-20, 21-40, etc.

Reliability The measure of consistency for assessment instruments. A reliable test will yield similar scores when abilities or knowledge are similar.

Rubric Refers to a set of scoring guidelines or standards to describe levels of student achievement on performance tests. Answers the question: What does mastery (and varying degrees of mastery) at this task look like?

Sampling A way to get information about a large group by examining only a small number of the group (the sample), or by giving all members small segments of the test. When conducted properly, the results are considered highly reliable.

Skill competencies A concept, skill, or attitude that is essential to an occupation; the term is often used to refer to both skill competencies and skill standards.

Skill standard The level of attainment or performance established for a skill competency.

Standard Statement of expected accomplishment; identified levels of accomplishment and/or performance of specific criteria.

Standardized Test An assessment instrument given to a large number of persons under similar conditions in order to yield comparable scores; includes national norm-referenced tests designed by publishers, such as ITBS, ACT, SAT.

Stanine Stanines, like percentile ranks, indicate a student's relative standing in a norm group. Stanines are normalized standard scores that range from a low of 1 to a high of 9, with 5 designating average performance.

Teacher-developed Test An assessment tool designed by the classroom teacher to check student's understanding; may be norm-referenced (designed to measure differences among the individuals in the class), criterion-referenced (specifying minimum levels of acceptable performance on specific objectives), and/or performance-based (demonstration of a specific skill or task). Since these tests are generally not identical from classroom to classroom or from school to school, this type of test cannot be used to compare students in separate locations.

Validity The measure of accuracy for assessment instruments. A valid test measures what we want it to measure, rather than extraneous variables. It also refers to the reliability of the process through which a test was developed.

Vocational program A sequence of courses designed to prepare students for an occupation or occupation area that typically requires education below the baccalaureate level.

Glossary

Harcourt Brace: Glossary of Measurement Terms:

BASIC MEASUREMENT CONCEPTS

Ability: A characteristic indicative of an individual's competence in a particular field. The word "ability" is frequently used interchangeably with aptitude, although many psychologists use "ability" to include what others term "aptitude" and "achievement." (See [Aptitude](#).)

Academic Aptitude (See [Scholastic Aptitude](#).)

Achievement/Ability Comparison (AAC): The relationship between an individual's score on a subtest of the *Stanford Achievement Test Series* or the *Metropolitan Achievement Tests* and the scores of other students of similar ability as measured by the *Otis-Lennon School Ability Test*. If a student's achievement test score is higher than those of students of similar ability, the AAC is HIGH. If the achievement score is about the same as the scores of similar-ability students, the AAC is MIDDLE; if the score is lower, the AAC is LOW.

Age Norms: The distribution of test scores by age of test takers. For example, a norms table may be provided for 9 year olds. This age-norms table would present such information as the percentage of 9 year olds who score below each raw score on the test. (See [Norms](#).)

Anecdotal Data: Data obtained from a written description of a specific incident in an individual's behavior (an anecdotal record). The written report should be an objective account of behavior considered significant for the understanding of the individual.

Aptitude: A combination of characteristics, whether native or acquired, that are indicative of an individual's ability to learn or to develop proficiency in some particular area if appropriate education or training is provided. Aptitude tests include those of general academic (scholastic) ability; those of special abilities, such as verbal, numerical, mechanical, or musical; tests assessing "readiness" for learning; and tests that measure both ability and previous learning, and are used to predict future performance—usually in a specific field, such as foreign language, shorthand, or nursing.

Calibrated Difficulty Level: A scale value that expresses how difficult a test item is. This value differs from the conventional difficulty index. (See [Difficulty Index](#).) The origin of the scale is arbitrary, but the lower the value, the easier the item.

Construct Validity (See [Validity](#).)

Content Validity (See [Validity](#).)

Correlation: The degree of relationship between two sets of scores. A correlation of 0.00 denotes a complete absence of relationship. A correlation of plus or minus 1.00 indicates a perfect (positive or negative) relationship. Correlation coefficients are used in estimating test reliability and validity.

Criterion-Referenced (Content-Referenced) Test: Terms often used to describe tests that are designed to provide information about the specific knowledge or skills possessed by a student. Such tests usually cover relatively small units of content and are closely related to

instruction. Their scores have meaning in terms of what the student knows or can do, rather than in (or in addition to) their relation to the scores made by some norm group. Frequently, the meaning is given in terms of a cutoff score, for which people who score above that point are considered to have scored adequately ("mastered" the material), while those who score below it are thought to have inadequate scores.

Criterion-Related Validity (See [Validity](#).)

Cumulative Percent (See [Percentile Rank](#).)

Deviation IQ (DIQ): An age-based index of general mental ability. It is based on the difference between a person's score and the average score for persons of the same chronological age. Deviation IQ scores from most current scholastic aptitude tests are standard scores with a mean of 100 and a standard deviation of 15 or 16 for each defined age group. Thus, the DIQ is a transformed score equal to 15 (or 16) $z + 100$. (See z -score and Standard Score.) Some people are moving away from calling such a score on a mental or scholastic ability test an IQ. The *Otis-Lennon School Ability Test*, for example, reports a School Ability Index. (See [School Ability Index](#).)

Deviation Score (x): The score for an individual minus the mean score for the group; i.e., the amount a person deviates from the mean ($x = X - \bar{X}$).

Diagnostic Test: A test used to "diagnose" or analyze; that is, to locate an individual's specific areas of weakness or strength, to determine the nature of his or her weaknesses or deficiencies, and, if possible, to suggest their cause. Such a test yields measures of the components or subparts of some larger body of information or skill. Diagnostic achievement tests are most commonly prepared for the skill subjects.

Difference Score: Difference between two scores for the same individual.

Difference Score Reliability: Reliability of the distribution of differences between two sets of scores. These scores could be on two different subtests, or on a pre- and posttest, where the difference score is typically called a gain score. The meaning of the term "reliability" is the same for a set of difference scores as for a distribution of regular test scores. (See [Reliability](#).) However, since difference scores are derived from two somewhat unreliable scores, difference scores are often quite unreliable. This must be kept in mind when interpreting difference scores.

Difficulty Index: The percent of students who answer an item correctly, designated as p . (At times defined as the percent who respond incorrectly, designated as q .)

Discrimination Index: The extent to which an item differentiates between high-scoring and low-scoring examinees. Discrimination indices generally can range from -1.00 to +1.00. Other things being equal, the higher the discrimination index, the better the item is considered to be. Items with *negative* discrimination indices are generally items in need of rewriting.

Distracters: An incorrect choice in a multiple-choice or matching item (also called a foil).

Equivalent Forms: Any of two or more forms of a test that are closely parallel with respect to content and the number and difficulty of the items included. Equivalent forms should also yield very similar average scores and measures of variability for a given group. Also called parallel or alternate forms.

Error of Measurement: The amount by which the score actually received (an observed score) differs from a hypothetical true score. (See also [Standard Error of Measurement](#).)

Frequency: The number of times a given score (or a set of scores in an interval grouping) occurs in a distribution.

Frequency Distribution: A tabulation of scores from low to high or high to low showing the number of individuals who obtain each score or fall within each score interval.

Gain Score: Difference between a posttest score and a pretest score.

Grade Equivalent (G.E.): A norm-referenced score; the grade and month of the school year for which a given score is the actual or estimated average. A grade equivalent is based on a 10-month school year. If a student scores at the average of all fifth graders tested in the first month of the school year, he/she would obtain a G.E. of 5.1. If the score was the same as the average for all fifth graders tested in the eighth month, the grade equivalent would be 5.8. There are some problems with the use of grade equivalents, and caution should be used when interpreting this type of score. For example, if a student at the end of fourth grade obtains a G.E. of 8.8 on a math subtest, this does *not* mean that the child can do eighth-grade work. Rather, it means that the child obtained the same score as an average student in the eighth month of the eighth grade, had the eighth-grade student taken the fourth-grade test.

Grade Norms: The distribution of test scores by the grade of the test takers. (See [Age Norms](#) and [Norms](#).)

Item Analysis: The process of examining students' responses to test items to judge the quality of each item. The difficulty and discrimination indices are frequently used in this process. (See [Difficulty Index](#) and [Discrimination Index](#).)

Item Difficulty: (See [Difficulty Index](#).)

Item Discrimination: (See [Discrimination Index](#).)

Latent-Trait Scale: A scaled score obtained through one of several mathematical approaches collectively known as Latent-Trait procedures or Item Response Theory. The particular numerical values used in the scale are arbitrary, but higher scores indicate more knowledgeable people or more difficult items. (See [Scaled Score](#).)

Local Percentile (See Percentile)

Mastery Level: The cutoff score on a criterion-referenced or mastery test. People who score at or above the cutoff score are considered to have mastered the material; people who score below the cutoff score are considered to be nonmasters. "Mastery" in this sense is an arbitrary judgment. A cutoff score can be determined by several different methods. Each method often results in a different cutoff score.

Mastery Test: A test designed to determine whether a student has mastered a given unit of instruction or a single knowledge or skill; a test giving information on what a student knows, rather than on how his or her performance relates to that of some norm group.

Mean (\bar{X}): The arithmetic average of a set of scores. It is found by adding all the scores in the distribution and dividing by the total number of scores.

Median (Md): The middle score in a distribution or set of ranked scores; the point (score) that divides a group into two equal parts; the 50th percentile. Half the scores are below the

median, and half are above it.

Mode: The score or value that occurs most frequently in a distribution.

N: The symbol commonly used to represent the number of cases in a group.

National Percentile (See [Percentile](#).)

Normal Curve Equivalents (NCEs): Normalized standard scores with a mean of 50 and a standard deviation of 21.06. (See Standard Score.) The standard deviation of 21.06 was chosen so that NCEs of 1 and 99 are equivalent to percentiles of 1 and 99. There are approximately 11 NCEs to each stanine. (See [Stanines](#).)

Normal Distribution: A distribution of scores or other measures that in graphic form has a distinctive bell-shaped appearance. In a normal distribution, the measures are distributed symmetrically about the mean. Cases are concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. The assumption that many mental and psychological characteristics are distributed normally has been very useful in test development work.

Figure 1 below is a normal distribution. It shows the percentage of cases between different scores as expressed in standard deviation units. For example, about 34% of the scores fall between the mean and one standard deviation above the mean.

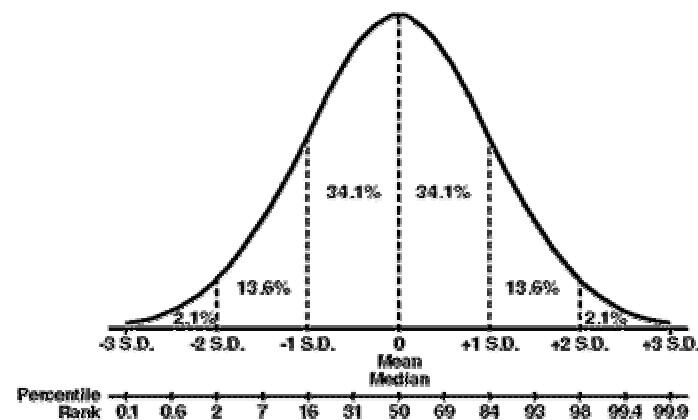


Figure 1. A Normal Distribution.

Norms: The distribution of test scores of some specified group called the norm group. For example, this may be a national sample of all fourth graders, a national sample of all fourth-grade males, or perhaps all fourth graders in some local district.

Norms vs. Standards: Norms are not standards. Norms are indicators of what students of similar characteristics did when confronted with the same test items as those taken by students in the norms group. Standards, on the other hand, are arbitrary judgments of what students *should* be able to do, given a set of test items.

Norm-Referenced Test: Any test in which the score acquires additional meaning by comparing it to the scores of people in an identified norm group. A test can be both norm- and criterion-referenced. Most standardized achievement tests are referred to as norm-referenced.

Objectives: Stated, desirable outcomes of education.

Out-of-Level Testing: The activity of administering a test level that is different from the one designated for a student of a particular age or in a particular grade. For example, a fourth grader might be given a test level designated for use in Grade 2. Out-of-level testing is used so that students can be tested on the content appropriate to their current level of functioning; that is, above or below their grade placement or age.

p-Value: The proportion of people in an identified norm group who answer a test item correctly; usually referred to as the difficulty index. (See [Difficulty Index](#).)

Percentile: A point on the norms distribution below which a certain percentage of the scores fall. For example, if 70% of the scores fall below a raw score of 56, then the score of 56 is at the 70th percentile. The term "local percentile" indicates that the norm group is obtained locally. The term "national percentile" indicates that the norm group represents a national group.

Percentile Band: An interpretation of a test score that takes into account measurement error. These bands, which are most useful in portraying significant differences between subtests in battery profiles, most often represent the range from one standard error of measurement below the obtained score to one standard error of measurement above it. For example, if a student had a raw score of 35, and if the standard error of measurement were 5, the percentile rank for a score of 30 to the percentile rank for a score of 40 would be the percentile band. We would be 68% confident the student's true percentile rank falls within this band. (See [Standard Error of Measurement](#) and [True Score](#).)

Percentile Rank: The percentage of scores falling below a certain point on a score distribution. (Percentile and percentile rank are sometimes used interchangeably.)

Profile: A graphic presentation of several scores expressed in comparable units of measurement for an individual or a group. This method of presentation permits easy identification of relative strengths or weaknesses across different tests or subtests.

Quartile: One of three points that divided the scores in a distribution into four groups of equal size. The first quartile [equation], or 25th percentile, separates the lowest fourth of the group; the middle quartile [equation], the 50th percentile or median, divides the second fourth of the cases from the third; and the third quartile [equation], the 75th percentile, separates the top quarter.

Raw Score: A person's observed score on a test, i.e., the number correct. While raw scores do have some usefulness, they should *not* be used to make comparisons between performance on different tests, unless other information about the characteristics of the tests is known. For example, if a student answered 24 items correctly on a reading test, and 40 items correctly on a mathematics test, we should not assume that he or she did better on the mathematics test than on the reading measure. Perhaps the reading test consisted of 35 items and the arithmetic test consisted of 80 items. Given this additional information we might conclude that the student did better on the reading test (24/35 as compared with 40/80). How well did the student do in relation to other students who took the test in reading? We cannot address this question until we know how well the class as a whole did on the reading test. Twenty-four items answered correctly is impressive, but if the average (mean) score attained by the class was 33, the student's score of 24 takes on a different meaning.

Readiness Test: A measure of the extent to which an individual has achieved the degree of maturity, or has acquired certain skills or information, needed to undertake some new learning activity successfully. For example, a reading readiness test indicates whether a child

has reached a developmental stage at which he may profitably begin formal reading instruction.

Regression Effect: Tendency of a posttest score (or a predicted score) to be closer to the mean of its distribution than the pretest score is to the mean of its distribution. Because of the effects of regression, students obtaining extremely high or extremely low scores on a pretest tend to obtain less extreme scores on a second administration of the same test (or on some predicted measure).

Reliability: The extent to which test scores are consistent; the degree to which the test scores are dependable or relatively free from random errors of measurement. Reliability is usually expressed in the form of a reliability coefficient or as the standard error of measurement derived from it. The reliability of a major classroom achievement test should be at least .60. The reliability of a standardized achievement or aptitude test should be at least .85. The higher the reliability coefficient the better, because this means there are smaller random errors in the scores. A test (or a set of test scores) with a reliability of 1.00 would have a standard error of zero and thus be perfectly reliable. (See [Standard Error of Measurement](#).)

Reliability Coefficients: Estimated by correlation between scores on two equivalent forms of a test, by the correlation between scores on two administrations of the same test, or through procedures known as internal-consistency estimates. Each of the three estimates pertains to a different aspect of reliability. One of the easier and more commonly used (by teachers) estimates of reliability is known as the Kuder-Richardson Formula #21 estimate. The formula is

$$KR_{21} = r_{xx} = \frac{n}{n-1} \left(1 - \frac{\overline{Xn} - \overline{X}}{ns_x^2} \right)$$

where n = number of items in the test

\overline{X} = mean of the test

s_x^2 = variance of the test

Reliability of Difference Scores (See [Difference Score Reliability](#).)

Scaled Score: A mathematical transformation of a raw score. Scaled scores are useful when comparing test results over time. Most standardized achievement test batteries provide scaled scores for such purposes. Several different methods of scaling exist, but each is intended to provide a continuous score scale across the different forms and levels of a test series.

Scaled-Score Band: An individual's scaled score plus and minus one standard error of measurement on the scaled-score metric. We can be 68% confident that the person's true scaled score is between the two end points of this band. (See [Standard Error of Measurement](#) and [True Score](#).)

Scholastic Aptitude: The combination of native and acquired abilities that are needed for school learning; the likelihood of success in mastering academic work as estimated from measures of the necessary abilities.

School Ability Index (SAI): Obtained from the *Otis-Lennon School Ability Test*, normalized standard score with a mean of 100 and a standard deviation of 16. (See [Deviation IQ](#) and [Standard Score](#).) An individual who had a School Ability Index of 116 would be one standard deviation above the mean, for example. This person would be at the 84th percentile for his or her age group. (See [Normal Distribution](#).)

Standard Age Scores: Normalized standard scores provided for specified age groups on each battery of a test. Typically, standard age scores have a mean of 100 and a standard deviation of 15.

Standard Deviation (S.D.) A measure of the variability, or dispersion, of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. In a normal distribution of scores, 68.3% of the scores are within the range of one S.D. below the mean to one S.D. above the mean. Computation of the S.D. is based upon the square of the deviation of each score from the mean. One way of writing the formula is as follows:

$$S.D. = \sqrt{\frac{\sum (X - \bar{X})^2}{N}} \quad \text{where } X = \text{raw score}$$

\bar{X} = mean

N = number of students

(See [Normal Distribution](#). <http://www.hbem.com/library/nordis>)

Standard Error of Measurement (SEM): The amount an observed score is expected to fluctuate around the true score. For example, the obtained score will not differ by more than plus or minus one standard error from the true score about 68% of the time. About 95% of the time, the obtained score will differ by less than plus or minus two standard errors from the true score.

$$SEM = S.D. \sqrt{1 - r_{xx}}$$

where S.D. = standard deviation and

r_{xx} = estimated reliability

The SEM is frequently used to obtain an idea of the consistency of a person's score or to set a band around a score. Suppose a person scores 110 on a test where the S.D. = 20 and [equation] = .91. Then:

$$SEM = 20\sqrt{1-.91} = 20\sqrt{.09} = 20(.3) = 6$$

We would thus say we are 68% confident the person's true score was between (110-1 SEM) and (110+1 SEM) or between 104 and 116.

Standard Score: A general term referring to scores that have been "transformed" for reasons of convenience, comparability, ease of interpretation, etc. The basic type of standard score, known as a z-score, is an expression of the deviation of a score from the mean score of the group in relation to the standard deviation of the scores of the group. Most other standard scores are linear transformations of z-scores, with different means and standard

deviations. (See [z-Score](#).)

Standards (See Norms vs. Standards)

Stanines: Expressed as a nine-point normalized standard score scale with a mean of 5 and a standard deviation of 2. Only the integers 1 to 9 occur. The percentage of scores at each stanine is 4, 7, 12, 17, 20, 17, 12, 7, and 4, respectively. While stanines are popular, they are actually less informative than, say, percentiles. For example, for three students with percentiles of 39, 41, and 59, the first would receive a stanine of 4, and the next two stanines of 5. We would thus be misled into inferring that the latter two students were the same, and different from the first with respect to the characteristic measured, whereas in reality the first two individuals are essentially the same, and different from the third.

Sometimes, the first three stanines are interpreted as being "below average," the next three as "average," and the top three stanines as "above average." This can be quite misleading. Suppose twins, Joe and Jim, have percentiles of 22 and 24, respectively. Joe would have a stanine of 3 and be considered "below average" whereas Jim would have a stanine of 4 and be considered average.

T-Score: A standard score with a mean of 50 and a standard deviation of 10. Thus a T-score of 60 represents a score one standard deviation above the mean. T-scores are obtained by the following formula:

$$T = 10z + 50$$

True Score: A score entirely free of error; a hypothetical value that can never be obtained by testing, since a test score always involves some measurement error. A person's "true" score may be thought of as the average of an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testing. The standard deviation of this infinite number of scores is known as the standard error of measurement. (See [Standard Error of Measurement](#).)

Validity: The extent to which a test does the job for which it is intended. The term validity has different connotations for different types of tests and, therefore, different kinds of validity evidence are appropriate for each.

1. Content validity: For achievement tests, content validity is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes (domain) about which inferences are to be made.

Typically, but not always, we wish to make inferences about the degree to which students have learned the material in a course. In those cases, the question of content validity is a question of the match and balance between the test items and the course content. At other times we wish to make different inferences. For example, we may wish to know (make inferences about) how well a group of students can perform the basic arithmetic functions even though we have not been teaching them directly but have been teaching set theory, different number bases, exponents, etc. In such a case, the content validity of a test would be the degree to which the test questions represent a balanced and adequate sampling of the domain of "arithmetic functions." The match is always between the questions asked and the domain of behavior about which inferences are to be made.

2. Criterion-related validity: The extent to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) some criterion measure.

Predictive validity refers to the accuracy with which a test is indicative of performance on a future criterion measure, e.g., scores on an academic aptitude test administered in high school to grade-point averages over four years of college. Evidence of concurrent validity is obtained when no time interval has elapsed between the administration of the test being validated and collection of data. Concurrent validity might be obtained by administering concurrent measures of academic ability and achievement, by determining the relationship between a new test and one generally accepted as valid, or by determining the relationship between scores on a test and a less objective criterion measure.

3. Construct validity: The extent to which a test measures some relatively abstract psychological trait or construct; applicable in evaluating the validity of tests that have been constructed on the basis of an analysis of the trait and its manifestation.

Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of their constructs by the relationships between their scores and pertinent external data.

Variability: The spread of dispersion of test scores, most often expressed as a standard deviation. (See [Standard Deviation](#).)

Variance: The square of the standard deviation.

Weighting: The process of assigning different weights to different scores in making some final decision. To do weighting correctly, one must convert all scores to a common scale or metric. For example, we cannot average temperatures measured with both the Celsius and Fahrenheit scale until the temperatures from one scale are converted to the other scale. For educational data, we should first convert all data to a common scale such as a z-score, a T-score, or some other standard score. Then, to combine scores, we must determine how much weight to give each score. Weights are usually assigned subjectively, based on the importance and/or quality, e.g., reliability, of the data.

z-Score: A type of standard score with a mean of zero and a standard deviation of one. (See [Standard Score](#).)

$$z = \frac{\text{raw score } (X) - \text{mean } (\bar{X})}{\text{standard deviation } (S.D.)}$$

Thus, for example, if three individuals have z-scores of -0.5, 0, and 1.0, we know the first scored one-half a standard deviation below the mean, the second scored right at the mean, and the third scored one standard deviation above the mean. If the distribution were normal these z-scores would have percentiles of about 33, 50, and 84, respectively. (See [Normal Distribution](#).)

APPENDIX F:

Articles: [“Alignment of Standards and Assessments as an Accountability Criterion”](#)

and

[“Fundamental Assessment Principles For Teachers and School Administrators”](#)

<http://ericae.net/pare/getvn.asp?v=7&n=21>

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2001, EdResearch.org.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

La Marca, Paul M. (2001). Alignment of standards and assessments as an accountability criterion. *Practical Assessment, Research & Evaluation*, 7(21). Available online: <http://edresearch.org/pare/getvn.asp?v=7&n=21>. This paper has been viewed 10,767 times since 9/17/01.

Alignment of Standards And Assessments as an Accountability Criterion

[Paul M. La Marca](#)
Nevada Department of Education

▶ Find similar papers in

[ERICA Full Text Library](#)
[Pract Assess, Res & Eval](#)
[ERIC RIE & CIJE 1990-](#)
[ERIC On-Demand Docs](#)

▶ Find articles in ERIC written by

[La Marca, Paul M.](#)

To make defensible accountability decisions based in part on student and school-level academic achievement, states must employ assessments that are aligned to their academic standards. Federal legislation and Title I regulations recognize the importance of alignment, which constitutes just one of several criteria for sound assessment and accountability systems. However, this seemingly simplistic requirement grows increasingly complex as its role in the test validation process is examined.

This paper provides an overview of the concept of alignment and the role it plays in assessment and accountability systems. Some discussion of methodological issues affecting the study of alignment is offered. The relationship between alignment and test score interpretation is also explored.

The Concept of Alignment

Alignment refers to the degree of match between test content and the subject area content identified through state academic standards. Given the breadth and depth of typical state standards, it is highly unlikely that a single test can achieve a desirable degree of match. This fact provides part of the rationale for using multiple accountability measures and also points to the need to study the degree of match or alignment both at the test level and at the system level. Although some degree of match should be provided by each individual test, complementary multiple measures can provide the necessary degree of coverage for systems alignment. This is the greater accountability issue.

Based on a review of literature (La Marca, Redfield, & Winter 2000), several dimensions of alignment have been identified. The two overarching dimensions are content match and depth match. Content match can be further refined into an analysis of broad content coverage, range of coverage, and balance of coverage. Both content and depth match are predicated on item-level comparisons to standards.

Broad content match, labeled categorical congruence by Webb (1997), refers to alignment at the broad standard level. For example, a general writing standard may indicate that "students write a variety of texts that inform, persuade, describe, evaluate, or tell a story and are appropriate to purpose and audience " (Nevada Department of Education, 2001 p. 14). Obviously this standard covers a lot of ground and many specific indicators of progress or objectives contribute to attainment of this broadly defined skill. However, item/task match at the broad standard level can drive the determination of categorical congruence with little consideration to the specific objectives being measured.

As suggested above, the breadth of most content standards is further refined by the specification of indicators or objectives. Range of coverage refers to how well items match the more detailed objectives. For example, the Nevada writing standard noted above includes a variety of specific indicators: information, narration, literary analysis, summary, and persuasion. Range of coverage would require measurement to be spread across the indicators. Similarly, the balance of coverage at the objective level should be judged based on a match between emphasis in test content and emphasis prescribed in standards documents.

Depth alignment refers to the match between the cognitive complexity of the knowledge/skill prescribed by the standards and the cognitive complexity required by the assessment item/task (Webb 1997, 1999). Building on the writing example, although indirect measures of writing, such as editing tasks, may provide some subject-area content coverage, the writing standard appears to prescribe a level of cognitive complexity that requires a direct assessment of writing to provide adequate depth alignment.

Alignment can best be achieved through sound standards and assessment development activities. As standards are developed, the issue of how achievement will be measured should be a constant consideration. Certainly the development of assessments designed to measure

Dimensions of Alignment

Content Match. *How well does test content match subject area content identified through state academic standards?*

- Broad content coverage. *Does test content address the broad academic standards? Is there categorical congruence?*
- Range of coverage. *Do test items address the specific objectives related to each standard?*
- Balance of coverage. *Do test items reflect the major emphases and priorities of the academic standards?*

Depth Match. *How well do test items match the knowledge and skills specified in the state standards in terms of cognitive complexity? A test that emphasized simple recall, for example, would not be well-aligned with a standard calling for students to be able to demonstrate a skill.*

expectations should be driven by academic standards through development of test blueprints and item specifications. Items/tasks can then be designed to measure specific objectives. After assessments are developed, a post hoc review of alignment should be conducted. This step is important where standards-based custom assessments are used and absolutely essential when states choose to use assessment products not specifically designed to measure their state standards. Whenever assessments are modified or passing scores are changed, another alignment review should be undertaken.

Methodological Consideration

An objective analysis of alignment as tests are adopted, built, or revised ought to be conducted on an ongoing basis. As will be argued later, this is a critical step in establishing evidence of the validity of test score or performance interpretation.

Although a variety of methodologies are available (Webb, 1999; Schmidt, 1999), the analysis of alignment requires a two-step process:

- a systematic review of standards and
- a systematic review of test items/tasks.

This two-step process is critical when considering the judgment of depth alignment. Individuals with expertise in both subject area content and assessment should conduct the review of standards and assessments. Reviewers should provide an independent or unbiased analysis; therefore, they should probably not have been heavily involved in the development of either the standards or the assessment items.

The review of standards and assessment items/tasks can occur using an iterative process, but Webb (1997, 1999) suggests that the review of standards precede any item/task review. An analysis of the degree of cognitive complexity prescribed by the standards is a critical step in this process. The subsequent review of test items/tasks will involve two decision points

- a determination of what objective, if any, an item measures and
- the items degree of cognitive complexity.

The subjective nature of this type of review requires a strong training component. For example, the concept of depth or cognitive complexity will likely vary from one reviewer to the next. In order to code consistently, reviewers will need to develop a shared definition of cognitive complexity. To assist in this process. Webb (1999) has built a rubric that defines the range of cognitive complexity.

Alignment Process

Conduct a systematic review of standards.

Conduct a systemic review of test items/tasks:

- Determine what objective(s) each item/task measures.
- Determine the degree of each item's cognitive complexity.

from simple recall to extended thinking. Making rubric training the first step in the formal evaluation process can help to reinforce the shared definition and ground the subsequent review of test items/tasks.

Systematic review of standards and items can yield judgments related to broad standard coverage, range of coverage, balance of coverage, and depth coverage. The specific decision rules employed for each alignment dimension are not hard and fast. Webb (1999) does provide a set of decision rules for judging alignment and further suggests that determination of alignment should be supported by evidence of score reliability.

Thus far the discussion has focused on the evaluation of alignment for a single test instrument. If the purpose of the exercise is ultimately to demonstrate systems alignment, the process can be repeated for each assessment instrument sequentially, or all assessment items/tasks can be reviewed simultaneously. The choice may be somewhat arbitrary.

However, there are advantages to judging alignment at both the instrument level and the system level. If, for example, decisions or interpretations are made based on a single test score, knowing the test's degree of alignment is critical. Moreover, as is typical of school accountability models, if multiple measures are combined prior to the decision-making or interpretive process, knowledge of overall systems alignment will be critical.

Why is alignment a key issue

In the current age of educational reform in which large-scale testing plays a prominent role, high-stakes decisions predicated on test performance are becoming increasingly common. As the decisions associated with test performance carry significant consequences (e.g., rewards and sanctions), the degree of confidence in, and the defensibility of, test score interpretations must be commensurably great. Stated differently, as large-scale assessment becomes more visible to the public, the roles of reliability and validity come to the fore.

Messick (1989) has convincingly argued that validity is not a quality of a test but concerns the inferences drawn from test scores or performance. This break from traditional conceptions of validity changes the focus from establishing different sorts of validity (e.g., content validity vs. construct validity) to establishing several lines of validity evidence, all contributing to the validation of test score inferences.

Alignment as discussed here is related to traditional conceptions of content validity. Messick (1989) states that "Content validity is based on professional judgments about the relevance of the test content to the content of a particular behavioral domain of interest and about the representativeness with which item or task content covers that domain" (p. 17). Arguably, the establishment of evidence of test relevance and representativeness of the target domain is a critical first step in validating test score interpretations. For example, if a test is designed to measure math achievement and a test score is judged relative to a set proficiency standard (i.e., a cut score), the interpretation of math proficiency will be heavily dependent on a match between test content and content area expectations.

Moreover, the establishment of evidence of content representativeness or alignment is intricately tied to evidence of construct validity. Although constructs are typically considered latent causal variables, their validation is often captured in measures of internal and external structure (Messick, 1989). Arguably the interpretation of measures of internal consistency and/or factor structures, as well as associations with external criterion, will be

informed by an analysis of range of content and balance of content coverage.

Therefore, alignment is a key issue in as much as it provides one avenue for establishing evidence for score interpretation. Validity is not a static quality, it is "an evolving property and validation is a continuing process" (Messick, p. 13). As argued earlier, evaluating alignment, like analyzing internal consistency, should occur regularly, taking its place in the cyclical process of assessment development and revision.

Discussion

Alignment should play a prominent role in effective accountability systems. It is not only a methodological requirement but also an ethical requirement. It would be a disservice to students and schools to judge achievement of academic expectations based on a poorly aligned system of assessment. Although it is easy to agree that we would not interpret a student's level of proficiency in social studies based on a math test score, interpreting math proficiency based on a math test score requires establishing through objective methods that the math test score is based on performance relative to skills that adequately represent our expectations for mathematical achievement.

There are several factors in addition to the subjective nature of expert judgments that can affect the objective evaluation of alignment. For example, test items/tasks often provide measurement of multiple content standards/objectives, and this may introduce error into expert judgments. Moreover, state standards differ markedly from one another in terms of specificity of academic expectations. Standards that reflect only general expectations tend to include limited information for defining the breadth of content and determining cognitive demand. Not only does this limit the ability to develop clearly aligned assessments, it is a barrier to the alignment review process. Standards that contain excessive detail also impede the development of assessments, making an acceptable degree of alignment difficult to achieve. In this case, prioritization or clear articulation of content emphasis will ease the burden of developing aligned assessments and accurately measuring the degree of alignment.

The systematic study of alignment on an ongoing basis is time-consuming and can be costly. Ultimately, however, the validity of test score interpretations depends in part on this sort of evidence. The benefits of confidence, fairness, and defensibility to students and schools outweigh the costs. The study of alignment is also empowering in as much as it provides critical information to be used in revising or refining assessments and academic standards.

References

- La Marca, P. M., Redfield, D., & Winter, P.C. (2000). *State Standards and State Assessment Systems: A Guide to Alignment*. Washington, DC: Council of Chief State School Officers.
- Messick, S. (1989). Validity. In R. L. Linn (Editor), *Educational Measurement (3rd Edition)*. New York: American Council on Education – Macmillan Publishing Company.
- Nevada Department of Education (2001). *Nevada English Language Arts: Content Standards for Kindergarten and Grades 1, 2, 3, 4, 5, 6, 7, 8 and 12*.
- Schmidt, W. (1999). Presentation in R. Blank (Moderator), The Alignment of Standards and Assessments. Annual National Conference on Large-Scale Assessment, Snowbird, UT.
- Webb, N. L. (1997). *Research Monograph No. 6: Criteria for Alignment of Expectations and*

Assessments in Mathematics and Science Education. Washington, DC: Council of Chief State School Officers.

Webb, N. L. (1999). *Alignment of Science and Mathematics Standards and Assessments in Four States*. Washington, DC: Council of Chief State School Officers.

The author would like to acknowledge Phoebe Winter, Council of Chief State School Officers, and Doris Redfield, Appalachia Educational Laboratory, for their assistance in critiquing this manuscript. I would like to acknowledge the CCSSO SCASS-CAS alignment work group for preliminary work in this area.

Correspondence concerning this article should be addressed to Paul M. La Marca, Director of Standards, Curricula, and Assessments, Nevada Department of Education, 700 E. Fifth St., Carson City, Nevada 89436. Electronic mail may be sent to plamarca@nsn.k12.nv.us.

Descriptors: Academic Standards; Educational Change; Evaluation Methods; Instructional Materials; *Item Analysis;

*Accountability; Achievement Gains

<http://ericae.net/pare/getvn.asp?v=7&n=8>

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2000, EdResearch.org.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

McMillan, James H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Evaluation*, 7(8). Available online: <http://edresearch.org/pare/getvn.asp?v=7&n=8>. This paper has been viewed 34,411 times since 9/23/00.

Fundamental Assessment Principles for Teachers and School Administrators

[James H. McMillan](#)

Virginia Commonwealth University

▶ Find similar papers in

[ERICA Full Text Library](#)
[Pract Assess, Res & Eval](#)
[ERIC RIE & CIJE 1990-](#)
[ERIC On-Demand Docs](#)

▶ Find articles in ERIC written by

[McMillan, James H.](#)

While several authors have argued that there are a number of "essential" assessment concepts, principles, techniques, and procedures that teachers and administrators need to know about (e.g. Calfee & Masuda, 1997; Cizek, 1997; Ebel, 1962; Farr & Griffin, 1973; Fleming & Chambers, 1983; Gullickson, 1985, 1986; Mayo, 1967; McMillan, 2001; Sanders & Vogel, 1993; Schafer, 1991; Stiggins & Conklin, 1992), there continues to be relatively little emphasis on assessment in the preparation of, or professional development of, teachers and administrators (Stiggins, 2000). In addition to the admonitions of many authors, there are established professional standards for assessment skills of teachers (*Standards for Teacher Competence in Educational Assessment of Students* (1990), a framework of assessment tasks for administrators (Impara & Plake, 1996), the Code of Professional Responsibilities in Educational Measurement (1995), the Code of Fair Testing Practices (1988), and the new edition of *Standards for Educational and Psychological Testing* (1999). If that isn't enough information, a project directed by Arlen Gullickson at The Evaluation Center of Western Michigan University will publish standards for evaluations of students in the near future.

The purpose of this article is to use suggestions and guidelines from these sources, in light of current assessment demands and contemporary theories of learning and motivation, to present eleven "basic principles" to guide the assessment training and professional development of teachers and administrators. That is, what is it about assessment, whether large-scale or classroom, that is fundamental for effective understanding and application? What are the "big ideas" that, when well understood and applied, will effectively guide good assessment practices, regardless of the grade level, subject matter, developer, or user of the results? As Jerome Bruner stated it many years ago in his classic, *The Process of Education*: "...the curriculum of a subject should be determined by the most fundamental

understanding that can be achieved of the underlying principles that give structure to that subject." (Bruner, 1960, p.31). What principles, in other words, provide the most essential, fundamental "structure" of assessment knowledge and skills that result in effective educational practices and improved student learning?

Assessment is inherently a process of professional judgment.

The first principle is that professional judgment is the foundation for assessment and, as such, is needed to properly understand and use all aspects of assessment. The measurement of student performance may seem "objective" with such practices as machine scoring and multiple-choice test items, but even these approaches are based on professional assumptions and values. Whether that judgment occurs in constructing test questions, scoring essays, creating rubrics, grading participation, combining scores, or interpreting standardized test scores, the essence of the process is making professional interpretations and decisions. Understanding this principle helps teachers and administrators realize the importance of their own judgments and those of others in evaluating the quality of assessment and the meaning of the results.

Assessment is based on separate but related principles of measurement evidence and evaluation.

It is important to understand the difference between measurement evidence (differentiating degrees of a trait by description or by assigning scores) and evaluation (interpretation of the description or scores). Essential measurement evidence skills include the ability to understand and interpret the meaning of descriptive statistical procedures, including variability, correlation, percentiles, standard scores, growth-scale scores, norming, and principles of combining scores for grading. A conceptual understanding of these techniques is needed (not necessarily knowing how to compute statistics) for such tasks as interpreting student strengths and weaknesses, reliability and validity evidence, grade determination, and making admissions decisions. Schafer (1991) has indicated that these concepts and techniques comprise part of an essential language for educators. They also provide a common basis for communication about "results," interpretation of evidence, and appropriate use of data. This is increasingly important given the pervasiveness of standards-based, high-stakes, large-scale assessments. Evaluation concerns merit and worth of the data as applied to a specific use or context. It involves what Shepard (2000) has described as the systematic analysis of evidence. Like students, teachers and administrators need analysis skills to effectively interpret evidence and make value judgments about the meaning of the results.

Assessment decision-making is influenced by a series of tensions.

Competing purposes, uses, and pressures result in tension for teachers and administrators as they make assessment-related decisions. For example, good teaching is characterized by assessments that motivate and engage students in ways that are consistent with their philosophies of teaching and learning and with theories of development, learning and motivation. Most teachers want to use constructed-response assessments because they

believe this kind of testing is best to ascertain student understanding. On the other hand, factors external to the classroom, such as mandated large-scale testing, promote different assessment strategies, such as using selected-response tests and providing practice in objective test-taking (McMillan & Nash, 2000). Further examples of tensions include the following.

- Learning vs auditing
- Formative (informal and ongoing) vs summative (formal and at the end)
- Criterion-referenced vs norm-referenced
- Value-added vs. absolute standards
- Traditional vs alternative
- Authentic vs contrived
- Speeded tests vs power tests
- Standardized tests vs classroom tests

These tensions suggest that decisions about assessment are best made with a full understanding of how different factors influence the nature of the assessment. Once all the alternatives understood, priorities need to be made; trade-offs are inevitable. With an appreciation of the tensions teachers and administrators will hopefully make better informed, better justified assessment decisions.

Assessment influences student motivation and learning.

Grant Wiggins (1998) has used the term 'educative assessment' to describe techniques and issues that educators should consider when they design and use assessments. His message is that the nature of assessment influences what is learned and the degree of meaningful engagement by students in the learning process. While Wiggins contends that assessments should be authentic, with feedback and opportunities for revision to improve rather than simply audit learning, the more general principle is understanding how different assessments affect students. Will students be more engaged if assessment tasks are problem-based? How do students study when they know the test consists of multiple-choice items? What is the nature of feedback, and when is it given to students? How does assessment affect student effort? Answers to such questions help teachers and administrators understand that assessment has powerful effects on motivation and learning. For example, recent research summarized by Black & Wiliam (1998) shows that student self-assessment skills, learned and applied as part of formative assessment, enhances student achievement.

Assessment contains error.

Teachers and administrators need to not only know that there is error in all classroom and standardized assessments, but also more specifically how reliability is determined and how much error is likely. With so much emphasis today on high-stakes testing for promotion, graduation, teacher and administrator accountability, and school accreditation, it is critical that all educators understand concepts like standard error of measurement. reliability

coefficients, confidence intervals, and standard setting. Two reliability principles deserve special attention. The first is that reliability refers to scores, not instruments. Second, teachers and administrators need to understand that, typically, error is underestimated. A recent paper by Rogosa (1999), effectively illustrates the concept of underestimation of error by showing in terms of percentile rank probable true score hit-rate and test-retest results.

Good assessment enhances instruction.

Just as assessment impacts student learning and motivation, it also influences the nature of instruction in the classroom. There has been considerable recent literature that has promoted assessment as something that is integrated with instruction, and not an activity that merely audits learning (Shepard, 2000). When assessment is integrated with instruction it informs teachers about what activities and assignments will be most useful, what level of teaching is most appropriate, and how summative assessments provide diagnostic information. For instance, during instruction activities informal, formative assessment helps teachers know when to move on, when to ask more questions, when to give more examples, and what responses to student questions are most appropriate. Standardized test scores, when used appropriately, help teachers understand student strengths and weaknesses to target further instruction.

Good assessment is valid.

Validity is a concept that needs to be fully understood. Like reliability, there are technical terms and issues associated with validity that are essential in helping teachers and administrators make reasonable and appropriate inferences from assessment results (e.g., types of validity evidence, validity generalization, construct underrepresentation, construct-irrelevant variance, and discriminant and convergent evidence). Of critical importance is the concept of evidence based on consequences, a new major validity category in the recently revised *Standards*. Both intended and unintended consequences of assessment need to be examined with appropriate evidence that supports particular arguments or points of view. Of equal importance is getting teachers and administrators to understand their role in gathering and interpreting validity evidence.

Good assessment is fair and ethical.

Arguably, the most important change in the recently published *Standards* is an entire new major section entitled "Fairness in Testing." The *Standards* presents four views of fairness: as absence of bias (e.g., offensiveness and unfair penalization), as equitable treatment, as equality in outcomes, and as opportunity to learn. It includes entire chapters on the rights and responsibilities of test takers, testing individuals of diverse linguistic backgrounds, and testing individuals with disabilities or special needs. Three additional areas are also important:

- Student knowledge of learning targets and the nature of the assessments prior to instruction (e.g., knowing what will be tested, how it will be graded, scoring criteria).

anchors, exemplars, and examples of performance).

- Student prerequisite knowledge and skills, including test-taking skills.
- Avoiding stereotypes.

Good assessments use multiple methods.

Assessment that is fair, leading to valid inferences with a minimum of error, is a series of measures that show student understanding through multiple methods. A complete picture of what students understand and can do is put together in pieces comprised by different approaches to assessment. While testing experts and testing companies stress that important decisions should not be made on the basis of a single test score, some educators at the local level, and some (many?) politicians at the state or the national level, seem determined to violate this principle. There is a need to understand the entire range of assessment techniques and methods, with the realization that each has limitations.

Good assessment is efficient and feasible.

Teachers and school administrators have limited time and resources. Consideration must be given to the efficiency of different approaches to assessment, balancing needs to implement methods required to provide a full understanding with the time needed to develop and implement the methods, and score results. Teacher skills and knowledge are important to consider, as well as the level of support and resources.

Good assessment appropriately incorporates technology.

As technology advances and teachers become more proficient in the use of technology, there will be increased opportunities for teachers and administrators to use computer-based techniques (e.g., item banks, electronic grading, computer-adapted testing, computer-based simulations), Internet resources, and more complex, detailed ways of reporting results. There is, however, a danger that technology will contribute to the mindless use of new resources, such as using items on-line developed by some companies without adequate evidence of reliability, validity, and fairness, and crunching numbers with software programs without sufficient thought about weighting, error, and averaging.

To summarize, what is most essential about assessment is understanding how general, fundamental assessment principles and ideas can be used to enhance student learning and teacher effectiveness. This will be achieved as teachers and administrators learn about conceptual and technical assessment concepts, methods, and procedures, for both large-scale and classroom assessments, and apply these fundamentals to instruction.

Notes:

An earlier version of this paper was presented at the Annual Meeting of the American Educational Research Association, New Orleans, April 24, 2000.

References

- Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139-148.
- Bruner, J. S. (1960). *The process of education*. NY: Vintage Books.
- Calfee, R. C., & Masuda, W. V. (1997). Classroom assessment as inquiry. In G. D. Phye (Ed.) *Handbook of classroom assessment: Learning, adjustment, and achievement*. NY: Academic Press.
- Cizek, G. J. (1997). Learning, achievement, and assessment: Constructs at a crossroads. In G. D. Phye (Ed.) *Handbook of classroom assessment: Learning, adjustment, and achievement*. NY: Academic Press.
- Code of fair testing practices in education* (1988). Washington, DC: Joint Committee on Testing Practices (American Psychological Association). Available <http://ericae.net/code.htm>
- Code of professional responsibilities in educational measurement* (1995). Washington, DC: National Council on Measurement in Education. Available <http://www.unl.edu/buros/article2.html>
- Ebel, R. L. (1962). Measurement and the teacher. *Educational Leadership*, 20, 20-24.
- Farr, R., & Griffin, M. (1973). Measurement gaps in teacher education. *Journal of Research and Development in Education*, 7(1), 19-28.
- Fleming, M., & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools*, San Francisco: Jossey-Bass.
- Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research*, 79(2), 96-100.
- Gullickson, A. R. (1996). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement*, 23(4), 347-354.
- Impara, J. C., & Plake, B. S. (1996). Professional development in student assessment for educational administrators. *Educational Measurement: Issues and Practice*, 15(2), 14-19.
- Mayo, S. T. (1967). Pre-service preparation of teachers in educational measurement. U.S. Department of Health, Education and Welfare. Washington, DC: Office of Education/Bureau of Research.
- McMillan, J. H. (2001). *Essential assessment concepts for teachers and administrators*. Thousand Oaks, CA: Corwin Publishing Company. Available <http://www.amazon.com>
- McMillan, J. H., & Nash, S. (2000). Teachers' classroom assessment and grading decision making. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans.
- Rogosa, D. (1999). How accurate are the STAR national percentile rank scores for individual students? - An interpretive guide. Palo Alto, CA: Stanford University.
- Sanders, J. R., & Vogel, S. R. (1993). The development of standards for teacher competence in educational assessment of students, in S. L. Wise (Ed.), *Teacher training in measurement and assessment skills*, Lincoln, NB: Burros Institute of Mental Measurements.
- Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice*, 10, (1), 3-6.
- Shepard, L. A. (2000). The role of assessment in a learning culture. Paper presented at the Annual Meeting of the American Educational Research Association. Available <http://www.aera.net/meeting/am2000/wrap/praddr01.htm>
- Standards for educational and psychological testing* (1999). Washington, DC: American

Educational Research Association, American Psychological Association, National Council on Measurement in Education.
Standards for teacher competence in educational assessment of students. (1990). American Federation of Teachers, National Council on Measurement in Education, National Education Association. Available: <http://www.unl.edu/buros/article3.html>
Stiggins, R. J. (2000). Classroom assessment: A history of neglect, a future of immense potential. Paper presented at the Annual Meeting of the American Educational Research Association.
Stiggins, R. J., & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany, NY: State University of New York Press, Albany.
Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass. Available <http://www.amazon.com>
Contact Information:

James H. McMillan
Box 842020
Virginia Commonwealth University
Richmond, VA 23284-2020

Phone: 804 828-1332, x553
Fax: 804-225-3554
jmcmilla@saturn.vcu.edu

Descriptors: *Standards; Professional Standards; Test Scores; Student Evaluation

APPENDIX G:

Articles: [“Authentic Assessment-Basic
Definitions and Perspectives”](#)
[“Authentic Assessment Tools”](#)
[“Academic Testing Test Design and Construction”](#)
[“Improving Your Test Questions”](#)
[“Item Writing Guidelines”](#)
[“Writing Multiple-Choice Test Items”](#)
[“More Multiple-choice Item Writing Do’s and Don’ts”](#)

Authentic Assessment— Basic Definitions and Perspectives

Rodney L. Custer
Illinois State University

[Table of Contents](#)

As a graduate student, I vividly recall the response to the question, “So, what are the latest trends in assessment?” The question was being posed to a leading expert in vocational assessment by another professional colleague. The setting was a morning cup of coffee and my interest was piqued. The answer was immediate and simple. Authentic assessment.

A decade has come and gone since that time and much has occurred, including *A Nation at Risk*, *Goals 2000*, *SCANS* (Secretary’s Commission on Achieving Necessary Skills), and more. Behaviorism has largely yielded to cognitivism, with associated interest in such things as constructivism, situated cognition, metacognition, and yes, authentic assessment.

Considerable work has been done over this past decade in the area of assessment. Around the nation, states have, with varying degrees of success, developed performance standards. In most quarters, there has been a genuine attempt to target higher-order thinking skills (e.g., critical thinking and problem solving) and to emphasize connections and synthesis over fact-based disciplinary content. Predictably, the results have been mixed, with concerns about such things as “learning the basics,” confusion about content, and concerns about assessment.

At the same time, much has changed. National curriculum standards, which have been developed for many of the disciplines (e.g., science, mathematics, geography, etc.), emphasize inquiry, problem solving, critical thinking, synthesis, and authentic contexts. Changes in assessment practices have also occurred. Most states and standards efforts are promoting the use of a performance component in addition to (or in lieu of) objective-based testing. At times, this has taken the form of constructed response items; in other cases, states and school systems have experimented with incorporating more extensive performance-based activities into the assessment process.

In many respects, this decade of intensive activity has served to validate much of what has been occurring for many years in vocational education. Consider emphases such as “hands on,” “lab-based,” coops, and internships. For years, considerable work has been invested in identifying competencies and subsequently molding them into behavioral objectives. Although some assessment remained focused on the testing of facts, there has also been a rather natural concern for observing (watching students while they do something) and evaluating the quality of completed tasks (i.e., judging projects against established

criteria). To some considerable extent, many of the practices that have been *typical* in vocational education have emerged as *alternative* in the larger academic community.

At the same time, activity in the larger academic community is informing vocational education and the two have been drawn more closely together. Vocational education research and practice are being informed by the insights of cognitive learning theory. Those from traditional academic areas are looking to vocational educators for help with authentic contexts and activities. And both are learning more about the complex interactions and connections between authentic learning and assessment.

This monograph was conceptualized as a kind of contemporary retrospective analysis. All of the authors have, in various ways, conducted our professional work in areas that we would have a difficult time defining as either *vocational* or *academic*. Actually, it has been both. Collectively, we have worked actively and in various ways with the National Science Foundation, national and state departments of education, and the National Research Council. We have provided leadership to national standards projects and have been active with the American Educational Research Association (AERA) and the Association for Career and Technical Education (ACTE, formerly the American Vocational Association). As such, we bring a rich and varied set of experiences and perspectives to this discussion of authentic assessment in vocational education. We like it that way and believe that this mix of experiences has enriched our thinking. Throughout the pages of this monograph, we have not attempted to restrict our vision to only those materials that are most applicable to vocational, career, or technical education. Rather, we have attempted to address the key issues from within our varied and mixed perspectives. Our sense is that this mirrors the best of what is occurring across education.

Basic Definitions

Before moving into an overview of the chapters, it will first be helpful to clarify some terminology related to assessment. Three commonly used terms are alternative, authentic, and performance assessment. Conceptually and in practice, these terms tend to describe similar things.

Alternative Assessment

Perhaps the least descriptive and useful is the term “alternative assessment.” As the term indicates, alternative assessments are essentially any assessment practices or tools that are different from traditional practice; more specifically, different from *paper-and-pencil tests*. A more informative approach is that taken by Neill (1997), associate director of the National Center for Fair and Open Testing. Neill has identified seven defining principles for *new* assessments developed by the National Forum on Assessment. These principles have received widespread support among educators and civil rights leaders, based on a desire for radical

Table of Contents

reconstruction of assessment practices as well as an emphasis on student learning as central to assessment reform. The seven principles endorsed by the forum are as follows:

1. The primary purpose of assessment is to improve student learning.
2. Assessment for other purposes supports student learning.
3. Assessment systems are fair to all students.
4. Professional collaboration and development support assessment.
5. The broad community participates in assessment development.
6. Communication about assessment is regular and clear.
7. Assessment systems are regularly reviewed and improved.

Actually, there are many different definitions offered for alternative assessment and no single definition prevails. According to Hamayan (1995), alternative assessment refers to procedures and techniques that can be used within the context of instruction and can be easily incorporated into the daily activities of the school or classroom. Huerta-Macias (1995) contrasts alternative assessments with traditional testing by placing the emphasis on integrating and producing rather than on recalling and reproducing. These authors also note that the main goal of alternative assessments is to gather evidence about how students are approaching, processing, and completing real-life tasks in a particular domain.

The term alternative assessment provides an umbrella for a variety of nontraditional assessment methods and techniques such as direct assessment, authentic assessment, and performance assessment (Butts 1997). However, given the growth and refinement that have occurred over the past decade, the term suffers from a lack of precision.

Authentic Assessment

Authentic assessments are essentially those that embed assessment in real-world contexts. Wiggins (1993) describes authentic assessment as tasks and procedures in which students are engaged in applying skills and knowledge to solve “real-world” problems, giving the tasks a sense of authenticity. He goes on to define authenticity as that which replicates the challenges and standards of performance typically facing writers, businesspeople, scientists, community leaders, designers, and technical workers. To design an authentic assessment activity, teachers must first decide what are the actual performances that they want students to be good at and then they must decide how they can frame learning experiences in a meaningful context that provides the connections between real world experiences and school-based ideas (Lund 1997).

A number of criteria have been used to define and describe authentic assessment. Among these are the following (Lund 1997; Wiggins 1993):

- Engaging and worthy problems or questions of importance to students,
- Replicas of or analogies to the kinds of problems faced by adult citizens and consumers or professionals in the field,

- Tasks that require the student to produce a high-quality product and/or performance,
- Transparent or demystified criteria or standards,
- Response-contingent challenges in which the effect of both process and product/performance determines the quality of the results,
- Emphasis on “higher-level” thinking and more complex learning,
- Evaluation of the essentials of performance against well-articulated performance standards often expressed as rubrics, and
- Assessments so firmly embedded in the curriculum that they are practically indistinguishable from instruction.

At a minimum, authentic assessments are those that require real-world applications of skills and knowledge that have meaning beyond the assessment activity (Archbald and Newmann 1988). However, a review of the criteria listed here shows that the concept also has been extended to include complex performances, creation of significant products, and accomplishment of complex tasks using higher-order cognitive skills.

Performance Assessment or Performance-Based Assessment

At the most basic level, performance assessment involves asking students to do something and then observing and rating the process and the finished product against predetermined criteria or a standard. As with other terms used to describe the various forms of assessment, other definitions of performance assessment tend to blur this distinctive meaning. For example, Herman (1999), associate director of the National Center for Research on Evaluation Standards and Student Testing, states that the “essence of performance assessments—whether in the form of open-ended questions, essays, experiments or portfolios—is that they ask students to create something of meaning” (online, n.p.). Herman continues by observing that good performance assessment involves complex thinking and/or problem solving, addresses important disciplinary content, invokes authentic or real-world applications, and uses tasks that are instructionally meaningful. Stated in this way, performance assessment sounds very much like authentic assessment.

In reality, the distinctions among terms are probably relatively small and probably insignificant. For our purposes in this monograph, we have chosen to use the term *authentic assessment*, since it tends to draw the boundary more broadly than performance assessment (authentic assessment typically involves some form of performance) and more precisely than alternative assessment (which typically includes everything but traditional testing).

Overview of the Monograph

The four chapters that comprise this work address distinctively different aspects of authentic assessment. In chapter one, John Schell discusses the theoretical underpinnings of authentic assessment. Whereas vocational education has a long history of behaviorist-oriented, competency-based education, authentic assessment has increasingly been informed by contemporary cognitive and sociological

learning theory. An important focus of the chapter is on the value of authentic learning and assessment practices as a mechanism for promoting learning transfer. In the second chapter, Brian McAlister provides a review and synthesis of what the research literature has to say about the value of authentic assessment. This “value question” has two important dimensions. First, the question is asked about the inherent value of authentic assessment as an approach to assessment. The second question has to do with the effectiveness of authentic assessment as a mechanism for enhancing and promoting student learning. Chapter three moves to the more pragmatic end of the continuum. After an initial discussion of three key concepts associated with authentic assessment (connecting, reflecting, and feedback), John Scott provides a comprehensive overview of the “tools” that are commonly used for authentic assessment. In the final chapter, Marie Hoepfl addresses one of the more perplexing issues associated with authentic assessment: the issues and challenges of using authentic practices for large-scale, high-stakes assessments.

We have enjoyed the discussions that led to the development of this monograph. We hope that you will enjoy it and that it will serve to extend your thinking about the nature of assessment in general and authentic assessment in particular.

Authentic Assessment Tools

John Scott
The University of Georgia

[Table of Contents](#)

Skillful and effective teachers require students to analyze and synthesize information, apply what they have learned, and demonstrate their understanding of material according to specified criteria. They have developed learning and assessment experiences to engage students and teach them how to “produce,” rather than simply “reproduce” knowledge (Burke 1992, p. 5). In these classrooms, the emphasis shifts from facts and isolated knowledge to active learning, where students work together to examine information and issues, solve problems, and communicate ideas. These shifts in emphasis are often accompanied by changes in assessment practices typified by involving students in authentic tasks, measuring a variety of outcomes, and involving students in self-assessment and reflection.

The focus of this chapter is on the “tools” used to conduct authentic assessment. It is important to preface this discussion by thinking about some key contextual issues. As anyone who has ever worked with tools of any kind knows, tools can be (and often are) misused. They are often used in ways and for purposes other than those for which they were designed. To press the analogy still further, most “tool boxes” contain a diverse selection of tools, each of which are selected and used for various purposes. Appropriate tool selection and use is a function of the knowledge and skill of the “tool user.” Much the same is true of authentic assessment. The toolbox is full of tools; but we must first think carefully about the various contexts and purposes for which they are used.

Connecting, Reflecting, and Feedback

There are three important aspects or concepts that should accompany any type of authentic assessment: connecting, reflecting, and feedback.

Connecting

Across the nation, considerable attention is being directed toward the reform of testing and assessment. Much of this thrust is designed to extend assessment beyond testing, with its emphasis on facts and fragments of information, to authentic methods of assessment. A key feature of many of these authentic strategies is that students are required to connect facts, concepts, and principles together in unique ways to solve problems or produce products. Cognitive research has challenged the belief that learning and learning transfer occur simply by accumulating and storing bits of information (Shepard 1989, p. 4). Contemporary learning theory holds that learners gain understanding as they draw on and extend previously learned knowledge, construct new knowledge, and develop their own cognitive maps (connecting diagrams) interconnecting facts,

concepts, and principles. Research indicates that information learned and assessed as a linear set of facts fails to yield the kinds of in-depth understanding needed to function in our modern society.

Glaser (1988) describes a number of different types of evidence collected through assessment. One of the most important of these is “coherence of knowledge.” Glaser goes on to observe that beginners’ knowledge is spotty and superficial, but as learning progresses, understanding becomes integrated and structured. Thus assessment should tap the connectedness of concepts and the student’s ability to access interrelated chunks.

Authentic assessments are almost always framed in the form of learning experiences. These experiences are typically sequenced from simple to complex and are progressive in nature. An important role of teacher-facilitators is to help students connect the knowledge and skills learned in previous tasks and then extend them to related or more complex tasks. Transfer of knowledge and skills is enhanced when students recognize the connectedness of learning. A number of authentic assessments such as graphic organizers, writing samples, and portfolios require students to connect (or synthesize) what they have learned to produce finished products. Many technical tasks presented in technology-based programs require students to connect their previous knowledge of mathematics, science, social studies, and English to solve problems and complete tasks and projects.

Reflecting

The range of available options for teachers wishing to improve student assessment extends beyond the cognitive and psychomotor domains to include assessment of attitudes and other affective behaviors. The key element here is to help students develop their self-awareness and reflective skills. Students need to learn how to assess their own work and to think about their thinking. A key aspect of many forms of authentic assessment is the opportunities that are provided for students to reflect on their thinking, practices, and learning. The technical term for this type of reflective process is metacognition.

Robin Fogarty (1994), in her excellent book ***The Mindful School: How to Teach for Metacognitive Reflection***, defines metacognition as a sense of awareness—“knowing what you know and what you don’t know” (p. viii). Barell (1992) extends Fogarty’s definition to include feelings, attitudes, and dispositions because thinking involves not only cognitive operations but also the dispositions to engage in cognitive activities.

Burke (1994) notes that metacognitive reflections provide students with opportunities to manage and assess their own thinking strategies. “Metacognition involves the monitoring and control of attitudes, such as students’ beliefs about themselves, the value of persistence, the nature of work, and their personal responsibilities in accomplishing a goal” (p. 96). These attitudes are fundamental to all tasks in varying degrees, whether academic or nonacademic. Teachers need to provide opportunities for students to engage in the kind of metacognitive moni-

toring where they reflect on “what we did well, what we would do differently next time, and whether or not we need help” (p. 96).

Numerous researchers (Barell 1992 Fogarty, Perkins, and Barell 1992; and Perkins and Salomon 1992) have explored the critical relationship between metacognition and learning transfer. Barell (1992) states that “in order to transfer knowledge of skills from one situation to another, we must be aware of them; metacognitive strategies are designed to help students become more aware” (p. 259). Fogarty, Perkins, and Barell (1992) define transfer as “learning something in one context and applying it in another” (p. ix).

In the constructivist view of learning, individuals absorb information and make sense of that information through metacognitive reflection. Reflection allows individuals to recognize the gaps that exist in their understanding. As gaps are recognized and become significant to students, they are motivated to locate, apply, and connect previous learning as well as to construct new knowledge.

Burke (1994) and Fogarty (1994), in their works on metacognition, detail a number of metacognitive strategies that can be used by classroom teachers. These include such techniques as Mrs. Potter’s Questions, KWL charts, PMI charts, transfer journals, wrap-around, reflection page, learning logs, seesaw thinking, pie in the face, stem sentences and many others.

- Mrs. Potter’s questions: What were you expected to do in this assignment? What did you do well? If you had to do this task over, what would you do differently? What help do you need from me?
- The KWL strategy consists of a three-column chart in which one column (K) is devoted to what I **Know**, the second (W) to what I **Want** to know, and the third (L) to what I **Learned** after finishing this lesson or assignment.
- The PMI strategy is similar to the KWL chart except the first column (P) is devoted to the **Plus** or favorable things found about a learning experience, the second (M) focuses on the **Minuses** or unfavorable finding, and the third (I) is devoted to what the student found **Interesting** about the learning experience.

Descriptions of other metacognitive strategies can be found in Burke’s and Fogarty’s books. It is very important to provide opportunity for learners to reflect on what has been learned as teachers rush to “cover the content in the textbook” and prepare learners to “pass the test.” Many learners are unaware of their thinking processes while they are learning and trying to create personal meaning out of some learning experience. When asked to describe what they initially thought about a topic, how they began to create personal understanding about some content, and what they would be able to do with this new knowledge or skill, they can’t describe how they went about it and usually reply “I don’t know how I did it, I just did.” Students who are taught how to reflect on learning by using metacognitive reflection strategies should be able to monitor, assess, and improve their own thinking and learning performance.

Feedback

Another important outcome of authentic assessment has to do with providing feedback to learners related to significant objectives. Wiggins (1993) notes that many teachers erroneously believe they are providing feedback with test scores and coded comments such as “good work,” “vague,” and “awkward.” What is wanted and needed by learners is user-friendly information about performance and how improvement can be made. Learners need information that will help them self-assess and self-correct so that assessment becomes integrated throughout the learning experience.

Wiggins (1993) draws a subtle, but important, distinction between guidance and feedback. Guidance gives direction whereas feedback tells one whether or not they are on course. Guidance is typically teacher initiated and tends to be prescriptive. By contrast, feedback actively involves and engages the learner. Frequently, the process is collaborative and reflective; the teacher and student become partners in the learning process. Figuratively, feedback techniques are those experiences that help students see themselves and their performance more clearly. Throughout the assessment process, students are provided with real-time information about the quality of their performance.

Wiggins (1993) notes that feedback is more like a running commentary rather than measurement. It enables learners to monitor their performance, thinking about whether or not they are on the right track without labeling or censoring their performance. From this feedback perspective, the emphasis shifts from “measurement” as an end goal to “assessment” as an ongoing and continuous process. To maximize the effect, feedback should occur while the performance is underway, not just after it is evaluated.

Mastery of complex, integrative learning activities extends well beyond simply responding to probing questions following performance. Rather, it involves continuous feedback throughout the process of solving complex problems. Successful performance requires concurrent feedback inherent in the task itself or in the context in which the task is performed that enables learners to self-assess and self-correct as accurately as possible. Optimally, feedback is best when it becomes an integral part of students’ own mental processes, when they learn how to assess themselves. Similar to other real-life situations, feedback is comprised of a complex set of external (family members, friends, co-workers, and supervisors) and internal messages (reflective and metacognitive thinking).

Self-Assessment

One of the more exciting, but underused, dimensions of authentic assessment is student self-assessment. Students want to know how they are doing *while* they are performing some tasks and, even more, they want to know how well they did when the task is completed. In traditional assessment, students must wait until post-performance tests have been graded for feedback. In alternative assessment

Table of Contents

classrooms, students are encouraged to engage in self-assessment and to collaborate with teachers to review performance and decide the next steps in the learning process.

One of the key aspects of student self-assessment has to do with criteria (or standards). These criteria come in different forms. In “self-referenced” assessment, learners evaluate performance in light of their own goals, desires, and previous attainments and thus become more cognizant of present performance as well as steps that must be taken to extend their learning. In this type of self-assessment, standards are embedded in the value system and inherent goals of students. In “standards-referenced” self-assessment, learners compare their own characteristics of performance against established standards or criteria.

Self-assessment abilities represent a critical workplace skill. In the workplace, individuals are continuously faced with situations in which they must assess situations, make decisions, and then evaluate the quality of those decisions. This type of authentic, formal self-assessment activity is rare in most public schools and universities. In most schools, students rarely have the opportunity to evaluate their own performance, because teachers have assumed the assessment role. Teachers who bemoan student apathy, lack of personal investment in their own education, willingness to settle for minimal performance, and even cheating may not realize that they are experiencing the results of teacher-vested assessment. What if students could be genuinely empowered to engage in meaningful self-assessment? What if the locus of authority in the assessment process were to be shifted from teacher to student, where the authority is shared? What if students had a real voice in developing and assessing their own learning?

At this point, it is important to acknowledge that this vision of self-assessment is contingent on such things as students’ developmental level, maturity, and previous educational experiences. Self-assessment techniques are not uniformly appropriate and will not always work. However, students who are given the opportunity to become more engaged in the learning process and in assessing their own progress often do respond with intelligence, responsibility, and determination after a learning period in which they develop assessment skills (Mabry 1999). For example, D’Urso (1996) reports the results of a study of second-grade students involved in their own assessment. She concludes that students’ sense of self improved, their work became more meaningful to them, they became protective of the knowledge they had gained, and they began to reflect on what they knew as well as on what they still needed to discover. They discovered their own “voice” and developed a deeper sense of self.

Strategies and Tools

We now turn our attention to the tools themselves. These tools must be carefully selected to provide opportunities for students to practice and perform meaningful tasks that are reflective of life outside of the classroom. Authentic assessment starts with the selection of meaningful learning tasks. These tasks need to be organized and structured so that they are contextualized, integrative,

metacognitive (require students to think about thinking), related to the curriculum taught, flexible (require multiple applications of knowledge and skills), open to self-assessment and peer assessment, contain specified standards and criteria, and are ongoing and formative (Weber 1999).

Mabry (1999) notes that we must match purpose or outcome expectations with assessment strategies. “What do we want to assess—and do we really need to assess it?” “Why do we want to assess it—what will we do with the results?” “How should we assess—how can we get the information we need?” “How can we assess without harmful side effects?” (p. 41). The central issue here has to do with “tool selection.” Given a particular problem, situation, or set of questions, teachers need to learn to ask, “What is the best tool for the job?”

Teachers will need to use a variety of assessment tools and techniques in order to enable all students to have a more complete picture of their growth and achievement. The National Center for Research in Vocational Education study *Using Alternative Assessment in Vocational Education* (Stecher et al. 1997) identified four categories of alternative assessment that are widely used in vocational education: (1) written assessments, including selected response types such as multiple choice and constructed responses types such as essay items or writing samples; (2) performance tasks; (3) senior projects including research papers, performance projects, and oral presentations; and (4) portfolios. With the development of computer-based simulation software, additional possibilities are being developed.

A wide variety of assessment tools are available to teachers and students. As one reviews the list of tools, it will become immediately obvious that there is scant distinction to be made between performance activities and assessment techniques. A key feature of authentic assessment is a “blurring” of the distinctions typically drawn between classroom activities and assessment (see Figure 1).

The kinds of performance activities shown in Figure 1 can serve as a basis for developing authentic assessments to transform assessment practices from summative and teacher directed to formative and student centered. A detailed discussion of each of these performance activities and how to structure assessment components is beyond the scope of this work. However, it is useful to make some general observations about the usefulness of these techniques as well as ideas for implementation. Following the general overviews, three performance activities (learning logs and journals, portfolios, and projects) are discussed in more detail. There is a growing body of well-illustrated resources available that are designed to help teachers structure authentic assessments. One particularly useful resource for authentic assessment tools is Skylight Professional Development < www.skylightedu.com > .

Graphic Organizers and Concept Mapping			Table of Contents
<ul style="list-style-type: none"> • Concept maps • Data tables • Cause and effect diagrams • Graphs • Run control charts • Flowcharts • Pareto diagrams 	<ul style="list-style-type: none"> • Correlation/scatter diagrams • Idea webs/graphic organizers • Geographic maps • Time lines • Venn diagrams 	<ul style="list-style-type: none"> • Event chains • Histograms • PMI strategy reports • Mrs. Potter's questions • Connecting elephants • Big idea generation • Ranking ladders • Mind maps 	
Performance Products			
<ul style="list-style-type: none"> • Business letters • Autobiographies • Editorials • Displays • Drawings/illustrations • Experiments • Essays • Surveys • Storyboard reports • Job applications • Book reviews • Bulletins • Critiques • Crossword puzzles • Designs • Requisitions 	<ul style="list-style-type: none"> • Vitas/Resumes • Inventions • Lab reports • Information-seeking letters • Management plans • Math problems • Geometry problems • Models • Writing samples • Job searches • Cartoons or comics • Collages • Consumer reports • Handbooks • Booklets • Home projects 	<ul style="list-style-type: none"> • Pamphlets • Observation reports • Research reports • Posters • Workplace scrapbooks • Grant applications • Team reports • Career plans • Video yearbooks • Training plans • Exhibits • Ballads • Announcements • Biographies • Questionnaires • Technical repairs 	
Live Performances and Presentations			
<ul style="list-style-type: none"> • Interviews • Issues/controversy • Workplace skits • Slide shows/video • Human graphs • Announcements 	<ul style="list-style-type: none"> • Games/quiz bowls • Student-led conferences • Story time/anecdotes • Prepared and extemporaneous speeches 	<ul style="list-style-type: none"> • Commercials • Demonstrations • Newscasts • Plays-TV/radio broadcasts 	

Figure 1. Authentic assessment tools/performance activities

Graphic Organizers and Concept Mapping

Graphic organizers are visual representations of mental maps using important skills such as sequencing, comparing, contrasting, and classifying. They involve students in active thinking about relationships and associations and help students make their thinking visible. Many students have trouble connecting or relating new information to prior knowledge because they cannot remember things. Graphic organizers help them remember because they make abstract ideas more visible and concrete. This is particularly true for visual learners who need graphic organizers to help them organize information and remember key concepts (Burke 1994).

Teachers can help students use graphic organizers by modeling and using topics that can be easily understood. Students can develop skills in developing graphic organizers if they are allowed to work first in small groups and can select a topic of their choice related to the lesson content.

Although graphic organizers are learning tools, they can also effectively be used as authentic assessment tools. Teachers who involve students with graphic organizers need to develop exemplary models that can be used for assessment. Criteria describing what content and relationships should be visually shown in student work need to be developed and used in rubric (scoring) form to make assessments more objective. Similar to essay questions, which require written expression in a connected manner, graphic organizers require students to present information in written and visual format. Graphic organizers also can be used as a test item format to assess student learning. This provides students with a creative and engaging way of expressing what they know and are able to do.

Performance Products

Many of the performance activities are end products of learning that can be assessed by rubrics (scoring forms) and other assessment tools designed to measure both processes and product quality.

Teachers who use authentic performance products provide students with opportunities to construct knowledge in real-world contexts so they can understand what they have learned. These products serve as a culminating experience in which students can retrieve previous learning, organize important information, and complete an assigned activity showing mastery of what they have learned.

Some teachers are reluctant to assign performance products because they do not feel comfortable grading them. They recognize that it takes time to construct exemplary models and to develop criteria and performance indicators required for rubric development. The key to assessing performance products is to set the standards and criteria in advance. Students who know the criteria that will be used to assess their work receive valuable instructional guidance in completing their products so they meet and/or exceed expectations.

As teachers recognize the importance of engaging students in making performance products, they will learn how to structure the learning environment to facilitate the process. They will also plan ahead to develop the tools needed to assess both the process of developing the product as well as the completed product. Scoring rubrics are one of the key assessment tools used for performance products. Information on how to construct and use them follows later.

Live Performances and Presentations

As with performance products, the key to effective assessment of live performances and presentations is establishing the criteria and performance indicators in advance. Criteria and performance indicators effectively organized into scoring

Table of Contents

rubrics provide examples of what students must do to demonstrate that they have learned at a specified level. The most important assessment strategy with live performances and presentations is to engage students in assessing their own performance first, followed by teacher assessment and an opportunity for students and teachers to interact over assessment findings. Live presentations involve two major assessment factors. One is the quality of the assigned work and the second is the demonstration of presentation skills. Scoring rubrics must include both of these factors.

Rubrics

Among the most common methods for student self-assessment are scoring rubrics. Marzano, Pickering, and McTighe (1993) have defined rubrics as “a fixed scale and list of characteristics describing performance for each of the points on the scale” (p. 10). Rubrics are scoring devices (or tools) that are designed to clarify, communicate, and assess performance. They are grading tools containing specific information about what is expected of students based on criteria that are often complex and subjective.

Rubrics typically contain two important features; they identify and clarify specific performance expectations and criteria, and they specify the various levels of student performance. In their simplest form, rubrics are checklists requiring a “yes” or “no” response. More complex rubrics include written standards of expected student performance with different levels of performance indicators describing student performance that meets or exceeds the standard.

There are as many different types of rubrics as there are rubric designers. Most rubrics fall under the two categories, holistic or analytical. Holistic rubrics consider performance as a totality, with the primary purpose being to obtain a global view of performance, typically on complex tasks or major projects. By contrast, analytical rubrics are designed to focus on more specific aspects of performance. Their purpose is to provide specific feedback on the level of performance on each major part, with the advantage of providing a detailed analysis of behavior or performance. These rubrics detect strengths and weaknesses and identify areas for refinement.

Rubrics of both types can be used appropriately for product and process assessment as well as for formative and summative assessment. It is also important to note that rubrics are typically developed and used as open communication devices. For example, it is not unusual for students to be involved in the process of developing the rubrics that will be used to assess their performance. Used in this way, rubrics become an effective mechanism for clarifying and openly communicating the expectations of learning activities. Many teachers share and discuss the contents of rubrics that will be used to assess an activity early in the process. As a result, the expectations are clarified and, in some cases, negotiated.

There are numerous advantages to using rubrics provide for both students and teachers:

- Enabling assessment to be more objective and consistent,
- Focusing attention of the assessor on the important outcomes with an assigned value for each,
- Demystifying the expectations for the student by assigning values for each expected outcome,
- Allowing students to identify strengths and to focus on weak areas while providing opportunity to revisit them,
- Prompting teachers to identify critical behaviors required for task completion and to establish the criteria for performance in specific terms,
- Encouraging students to develop a consciousness about the criteria they are to demonstrate in their performance as well as the criteria they can use to assess their own abilities and performance,
- Promoting an emphasis on formative as well as summative evaluation,
- Providing benchmarks against which to measure and document progress,
- Lowering student anxiety about what is expected of them,
- Ensuring that students' work is judged by the same standard, and
- Leading students toward high-quality performance.

There are some disadvantages as well. Rubrics can be time consuming to develop and use. Good rubrics also must be grounded in clearly identified and stated criteria or standards. In many cases, these have not yet been identified or developed. Once the criteria have been clarified, considerable work remains to clearly identify the key indicators that will be used to assess the various levels of attainment for each of the criteria. This is the hard work of solid, clear, and meaningful assessment. The expectations must be clarified and then the level of attainment must be described and clearly communicated.

Some general guidelines for involving students in constructing and using rubrics have been developed by Goodrich (1997):

1. Begin by looking at models. Show students examples of good and not so good work. Identify the characteristics that make the models good and the bad ones bad.
2. List the critical criteria for the performance. A good guide is to think about what you would need to include if you had to give feedback to a student who did poorly on a task. Students can be involved in discussing the models to begin a listing of what counts in high-quality work.
3. Articulate gradations of quality or determine the quality continuum. Describe the best and worst levels of quality, and then fill in the middle based on knowledge of common problems associated with the performance. Use descriptive terms such as Not yet, OK, and Awesome instead of failure, average, and excellent.
4. Engage students in using the rubrics created to evaluate the models given them in step 1 as practice in self-assessment and to pilot test the rubrics.

Table of Contents

5. Give students their task. As they work, stop them occasionally for self- and peer assessment using the rubrics provided.
6. Give students time to revise their work based on the feedback they received in step 5.
7. Use the same rubric students used to assess their work. This is made possible by including a scoring column for students, peers, and teachers.
8. Schedule a debriefing time with students to compare their rubric scoring with those completed by the teacher. Require students to reflect on the next steps in the learning process.

One excellent resource is ***Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model*** by Marzano, Pickering, and McTighe (1993), published by the Association for Supervision and Curriculum Development. This work contains many examples of rubrics for specific tasks and situations. Another approach to developing rubrics using a “shell” to cluster criteria according to valued workplace competencies (e.g., creative thinking, contributing citizen, problem solving, effective communication, etc.) was developed by Custer (1996).

Portfolios

Another alternative assessment tool that has attracted widespread popular attention is portfolios. Portfolios are collections of student work gathered over time. The contents of portfolios can range from comprehensive coverage containing a plethora of materials to those that are quite selective, containing only a limited number of student-selected items. Student portfolios offer a range of flexibility that makes the method attractive to a wide range of teachers and programs. The elements to be included in this type of assessment are almost endless. Several critical components of effective portfolios are—

- A thoughtful student-developed introduction to the portfolio,
- Reflection papers behind each major assignment of the portfolio,
- Scoring rubrics for portfolio entries that enable students to self-assess their work,
- Established models, standards, and criteria that enable students to select their best work to be included in the portfolio, and
- Student oral presentation of their portfolios to significant others such as peers, teachers, and parents.

Portfolio assessment offers many advantages, but Frazier and Paulson (1992) note that the primary value of portfolios is that they allow student the opportunity to evaluate their own work. Further, portfolio assessment offers students a way to take charge of their learning; it also encourages ownership, pride, and high self-esteem. Portfolios can be maintained over several years and can be used as “pass-

ports” as students move from one level of education to another. Portfolio passports can also be used as valuable tools for obtaining jobs in business and industry.

Portfolio assessment requires careful thought and preparation on the part of both teachers and students. Vavrus (1990) offers the following considerations and recommendations that should be considered in designing a portfolio assessment system.

- ***What will it look like?*** Portfolios must have both a physical structure (binder as well as the arrangement of documents within the portfolio) and a conceptual structure (underlying goals for student learning).
- ***What goes in?*** To answer this question, other questions must first be addressed: Who is the intended audience for the portfolios? What will this audience want to know about student learning? How will these audiences be involved in portfolio development? Will selected documents of the portfolio show aspects of student learning that traditional test results do not show? What kinds of evidence will best show student progress toward expected learning outcomes? Will the portfolio contain best works only, a progressive record of student growth, or both? Will the portfolio include more than finished pieces—for example, notes, ideas, sketches, drafts, and revisions?
- ***How will procedural and logistical issues be addressed?*** How will student working files and portfolios be kept secure? When will students select documents to include in their portfolios? When will some portfolio document be taken out to specialize the portfolio? What criteria or assistance will be provided to students so that they can reflect on their work, monitor their own progress, and select pieces for inclusion in the portfolio? Will students be required to provide a rationale or explanation for work selected for inclusion in the portfolio?
- ***How will portfolios be evaluated and who will be involved?*** It is critical that students be actively involved in assessing their own work. To facilitate student self-assessment teachers will have to answer some important questions. What factors will be evaluated such as achievement in relation to standards, student growth along a continuum, or both? What models, standards, criteria and instruments will have to be developed to guide assessment? When will portfolio entries be evaluated? Will other teachers be involved assessing portfolio elements? Will parents or guardians be involved in assessing the portfolio? If so, how?
- ***What will happen to the portfolio at the end of the semester or school year?*** Will they be turned over to students at the end of the course or school year to keep and use as they see fit? Will students be encouraged to keep their portfolios over an extended period of time and use them as “passports” for entry into other levels of education or to work?

Table of Contents

It is clear that portfolios are a way of collecting and packaging a comprehensive body of rich evaluation materials. The key is to think carefully through the many logistical, conceptual, and procedural issues that must be addressed in order for this tool to be used effectively. Portfolios should not be “a place to dump anything and everything” loosely related to a given course. Rather, their value as an assessment tool is maximized when they contain items that have been carefully and thoughtfully selected to address specified learning goals. At their best, portfolios can represent an extremely rich portrait of student ability and interest.

Learning Logs and Journals

Learning logs and journals are tools designed to cause students to reflect on what they have learned or are learning. Used properly, they encourage student self-assessment and provide a mechanism for making connections across the various subject matter areas. Journals have been used widely in English classes for many years. Now they are being adopted by other teachers to develop communication skills and to help students to make connections, examine complex ideas, and think about ways to apply what they have learned over an extended period of time. Herman, Aschbacher, and Winters (1992) indicated that the fundamental purpose of learning logs and journals is to “allow students to communicate directly with the teacher regarding individual progress, particular concerns, and reflections on the learning process” (p. 2).

A distinction can be made between learning logs and journals. Learning logs usually consist of short, objective entries under specific heading such as problem solving, observations, questions about content, lists of outside readings, homework assignments, or other categories designed to facilitate recordkeeping (Burke 1994). Student responses are typically brief, factual, and impersonal. Fogarty and Bellanca (1987) recommend teachers provide lead-ins or stem statements that encourage students responses that are analytical (breaking something down into its parts), synthetic (putting something together into a whole), and evaluative (forming judgment about the worth of something). Example log stems include the following: One thing I learned yesterday was..., One question I still have is..., One thing I found interesting was..., One application for this is..., and I need help with...

By contrast, journals typically include more extensive information and are usually written in narrative form. They are more subjective and focus more on feelings, reflections, opinions, and personal experiences. Journal entries are more descriptive, more spontaneous, and longer than logs. They are often used to respond to situations, describe events, reflect on personal experiences and feelings, connect what is being learned with past learning, and predict how what is being learned can be used in real life (Burke 1994). As with learning logs, stem statements can be used to help students target responses. Example lead-ins are as follows: My way of thinking about this is..., My initial observation is..., Upon reflection I...

Learning logs and journals can be used in the following ways (Burke 1994):

- Record key ideas from a lecture, video, presentation, field trip, or reading assignment,
- Make predictions about what will happen next in a story, video, experiment, event, situation, process, or lesson,
- Record questions and reflect on the information presented,
- Summarize main ideas of a lesson, article, paper, video, or speech,
- Connect the ideas presented to previous learning, or to other subjects or events in a person's life,
- Monitor change in an experiment or event over time,
- Brainstorm ideas about potential projects, papers, presentation, assignments, and problems,
- Help identify problems and record problem-solving techniques, or
- Track progress in solving problems, readings, homework assignments, projects, and experiences.

Learning logs and journals can be effective instructional tools to help students sharpen their thinking and communication skills. They give students the opportunity to interact with the teacher, lesson content, textbooks, and each other. They also afford students an opportunity to think about material, clarify confusion, discuss key ideas with others, connect with previous learning and experiences, and reflect on the personal meaning of subject matter. They provide a record over time of what has been presented and learned. Furthermore, logs and journals are typically best used to promote formative assessment, although they also can be structured to provide summative assessment information.

Projects

Many different types of projects can be developed to challenge students to *produce* something rather than *reproduce* knowledge on traditional tests. Projects allow students to demonstrate a variety of skills including communication, technical, interpersonal, organizational, problem-solving, and decision making skills (Burke 1994). Projects also provide students with opportunities to establish criteria for determining the quality of the planning and design processes, the construction process, and the quality of the completed project.

The Southern Regional Educational Board has published a guide to preparing a syllabus for its *High Schools that Work Program* that includes a major focus on projects as the centerpiece of curriculum, instruction, and evaluation. This guide, ***Designing Challenging Vocational Courses*** by Bottoms, Pucel, and Phillips (1997), describes the procedures required to select and sequence major course projects, develop project outlines, decide on an instructional delivery plan, and develop an assessment plan.

Several states, notably California and Kentucky, have made successful completion of a student-initiated culminating project (senior project) a part of their student assessment system. The California Department of Education (1994), in collabora-

Table of Contents

tion with the Far West Laboratory, has developed the Career-Technical Assessment Program (C-TAP), which includes a C-TAP project. The project is a major piece of “hands-on” work designed and completed by each student. The project becomes an instructional and assessment tool that allows students to demonstrate skills and knowledge learned in a sequenced instructional program. Completing the project provides a mechanism for students to plan, organize, and create a product or event. Through this process, students are able to pursue their own interests, meet professionals in the field who can offer advice and instruction related to their project, work cooperatively with others in certain parts of the project, and apply the knowledge and skills they have learned in other school subjects. Each student’s project must be related to the career-technical program in which they are enrolled and can take as little as a few weeks to complete or several months. Students are allowed to work on the project themselves or in small groups. There are four major sections of the C-TAP project:

1. *Plan*: A process that helps the student design the project
2. *Evidence of Progress*: Three pieces that show the student’s progress toward developing the final product
3. *Final product*: A final product that is the result of the student’s work
4. *Oral presentation*: An oral presentation in which the student describes the project, explains what skills were applied, and evaluates his or her work

C-TAP projects are evaluated in two ways with two separate scores being generated. First, the project is rated using a rubric focused on three evaluation dimensions: content, communication, and responsibility. Content pertains to career-technical knowledge and skills, communication relates to the overall presentation of work, and responsibility pertains to the student’s ability to complete work independently. The second score (also generated using a rubric) focuses on oral presentation skills including public speaking skills, content knowledge, and analysis. A student manual and a teacher guidebook contains the information necessary for the complete operation of the C-TAP program.

Summary

Many factors are driving assessment reform in this country, including an emphasis on constructivism and authenticity, standards, and higher-order thinking skills. These forces and others have stirred interest in the educational community to look for alternatives to traditional testing in order to give a more accurate and complete picture of student growth and achievement. Organizations that specialize in assessment (e.g., the Far West Laboratory and the Center for Research on Evaluation, Standards, and Student Testing) are working with school systems to develop and test alternative assessments. The preliminary results are quite promising in terms of reform in curriculum and instructional practice as well as increased student engagement in the learning and assessment process. Assessment of learning is truly a “work in process.” It is exciting to see the progress that has been made to move beyond teaching and testing fragmented lists of declarative knowledge in favor of involving students in applying knowledge in unique and authentic ways.

The challenge for teachers is to commit to change the way they teach and assess students as well as put forth the effort to develop and use alternative assessment strategies such as those described in this chapter. Every effort should be made to develop meaningful, authentic learning and assessment tasks that target the knowledge, skills, and attitudes necessary for learning and life. Educators must also learn how to organize and structure these tasks so that they are contextualized, integrative, flexible, and open to self-assessment and peer assessment. Additionally, a clear focus on standards and criteria must be maintained in a way that provides for both formative and summative procedures. Students should be encouraged to become actively involved in the assessment process through metacognitive reflection, establishing criteria and performance indicators required to develop effective scoring rubrics, and using these scoring instruments to assess their own work. Effective feedback is the key to improved student learning. Yet many teachers are reluctant to spend the time required to develop and exhibit exemplary models of expected performances and to teach students how to assess and regulate their own performance.

Considerable progress has been made in the 1990s in designing and implementing alternative assessments. There are many success stories that point toward systemic change in the way educators are structuring curriculum, delivering instruction, and assessing student growth and achievement. Much of this work closely mirrors work that has been done in vocational education for many years. The current shared interest between the vocational and academic communities holds promise for improving both as teachers share ideas, techniques, and tools across disciplines.

Authentic assessment supports change in curricula, teaching, and school organization. But the real question is “Do these new assessment methods and techniques contribute to improved student learning?” A growing number of teachers seem to think so. Reporting on the effects of authentic assessment in action at five schools, Darling-Hammond, Aneess, and Falk (1995) note that classroom interactions, student work, exhibitions, and hallway conversations provide widespread evidence of in-depth learning, intellectual habits of mind, high-quality products, and student responsiveness to rigorous standards.

http://www.uts.psu.edu/Test_construction_frame.htm**Penn State****University Testing Services****23 Willard Bldg.****814-863-2802****utest@psu.edu**

ACADEMIC TESTING TEST DESIGN AND CONSTRUCTION

In practice, the development and evaluation of any test to measure learning outcomes is a laborious but highly rewarding process requiring the collaborative effort of content and testing specialists. In this document we describe the methods used to construct high quality tests of academic knowledge and understanding for specific areas of academic instruction. We limit discussion to the traditional and very practical methods of test construction and evaluation associated with the Classical (or Weak True-Score) Test Theory.

The principles upon which effective test development rests are described in the documents "Academic Testing: [Classical Test Theory Approach](#)" and "Academic Testing: [Item Response Theory Approach](#)."

Measurement Plan

The process of test construction comprises a number of rigorous steps.

1. Specify the domain of knowledge and understanding, such as elementary linear algebra, that the student is required to learn.
2. Identify the mental tasks or processes that the student must use in dealing with the particular subject matter;
mental tasks such as recall, analysis, generalization, application, and discovery.
3. Write test questions or items that unambiguously assess the student's ability to deal with the knowledge domain
as described by Steps 1 and 2.
4. Administer the test items written in Step 3 to a random sample of the population for which the test is intended.
5. Analyze the item response data collected in Step 4. This step is commonly referred to as the item analysis step.
6. Based upon the results of Step 5, select and, if necessary, revise the "best" items.
7. Administer the items selected in Step 6 to another random sample of the target population.
8. Repeat Steps 5 through 7, inclusive, until an optimal pool of items is identified.

9. Transform the raw scores of the optimal test to some meaningful metric.

Test Item Preparation

The first three steps of the measurement plan are the sole responsibility of the content specialist, i.e., the teacher and his or her colleagues. In most cases, [Step 1](#) is rather straightforward. Not so for [Steps 2 and 3](#), however, particularly test-item writing. Despite being governed by common sense, the writing of "good" test items is a difficult and time-consuming task that is best characterized as an art. Gronlund (1982) provides a practical and readable introduction to the "art" of designing a test and writing test questions. Another excellent and highly recommended reference for test design and item writing is Haladyna (1999).

Item Format

An important consideration in test construction is the choice of how a particular test question should be asked and answered. A variety of item formats is available to the item writer even within the same test. Item formats for tests of learning outcomes fall into two broad classes: the free- or constructed-response format and the selected-response format.

The free- or constructed-response item format requires the student to supply the correct response. The most familiar free-response item format is the essay. Essay questions differ in how much freedom is permitted the student in making a response. "Restricted" essays require brief and precise answers to specific questions. "Extended" essays, on the other hand, reflect more comprehensive questions that allow greater freedom in structuring a response. A second type of free-response item format is the short-answer or completion item. The test item can be either a question, e.g., "If $x = 2$, then what does $2x + 5$ equal?" or a statement with a missing element to be provided by the student.

The selected-response item format requires the student to choose the correct response from within a set of possibly "correct" item responses or foils which accompanies the test item. The most common types of selected-response item formats are the true-false item, the multiple-choice item, and the matching item.

For the true-false item format the choice is between the truth and falsity of a statement, e.g., "John F. Kennedy was the 40th president of the United States." (False). The multiple-choice item format requires the student to select one of two or more possible answers to the question, e.g., "How many players make up the starting line-up of a major league baseball team?" {Seven, Eight, Nine, Ten, Eleven} (Nine). The matching item comprises two lists; a list of related questions, e.g., a set of quadratic equations, and a list of possible answers to the questions, e.g., a set of solutions to the equations. In this example the student's task is to "match" the solutions and equations.

Item Scoring

Scoring a test item can be either the assignment of a numerical value to a student's response or the placement of the response in one of two or more ordered categories.

With selected-response items—i.e., true-false, multiple-choice, and matching items—and the free-response short-answer or completion item, scoring typically involves assigning a "+1" to the correct response and a "0" to each incorrect response. In particular, the item score is a Bernoulli random variable S_i having $S_i = 1$ if item i is answered correctly and $S_i = 0$ if item i is answered incorrectly. A

total test score, X , is typically computed by summing the item scores over all items, i.e., $X = \sum_{i=1}^k S_i$ for a test of k items.

These item types are often referred to as objective items in the sense that personal judgment does not enter into the scoring. In most cases a computer program is used to score the items of the test.

A serious threat to the precision of a selected-response item score is guessing. When a student does not know the correct answer to a test question but guesses correctly, the item score for that student is biased. In general, the smaller the number of response options the greater the bias associated with guessing.

Restricted and extended essays, on the other hand, require expert assessment or evaluation of the student's response. The simplest method of essay scoring is holistic scoring which requires the expert "reader" to place the student's response in one of several ordered categories of response quality. More complex methods of essay scoring can be and are used. All methods of essay scoring are susceptible to scoring bias stemming from differences among readers in judging the quality of the same response. Scoring bias is an obvious threat to the precision and usefulness of the measure.

Item Analysis

How good is a test question or item in measuring a learning outcome? The answer depends upon the purpose of the test containing the item. A "good" test item is an item that is optimum relative to the purpose of the test, e.g., measuring knowledge or predicting success in a training program. We now describe a number of practical techniques for evaluating test items.

Item Difficulty

Item difficulty is defined as the probability of a correct response to the question, i.e., $\Pr\{S_i = 1\} = p_i$. The proportion of a random sample from the target population that correctly answers the test question (see [steps 4 and 5](#) of the Measurement Plan) is an estimator of p_i . For the i -th test item and a sample of size N ,

$$p_i = \frac{n_c}{N}$$

n_c is the number of individuals in the sample that correctly answers the question.

Items with p values close to zero (very "hard" items) or close to one (very "easy" items) do not, in general, contain much information about the performance of the target population, except for the tails of the distribution. For most applied measurement situations, tests having items with item difficulty values between 0.3 and 0.7, inclusive, will maximally differentiate students throughout the entire performance range. If, however, the intent is to discriminate at a specific level of performance—say for purposes of selection—then items with p values in the narrow interval consistent with that performance level are chosen.

Item Discrimination

Another important property of any test question is item discrimination. Item discrimination involves the strength of the relationship between a test item and the underlying (and unobservable) attribute being measured, e.g., knowledge or learning. Since the latent variable, Y , cannot be measured directly, we turn to a set of observable variables whose elements are related to Y . In most applications, this set turns out to be the set of item variables $\{s_i\}$. While it seems somewhat circular, the total score, X , is often taken as a measure of the underlying attribute, Y .

The stronger the relationship between the item score, s_i , and the total score, X , the greater the differentiation among examinees on Y due to the i -th item. The item score/total score point-biserial correlation, $r_{s_i, X}$, is a measure of item discrimination. In a sample of size N , the point-biserial correlation coefficient is

$$r_{s_i, X} = \frac{\bar{X}_i - \bar{X}}{s_X} \sqrt{\frac{p_i}{1 - p_i}}$$

In the formula, \bar{X}_i is the mean total score of the $N_i = p_i N$ examinees who correctly answered the item, \bar{X} is the mean total score of all the examinees, and s_X is the total score standard deviation. To avoid spurious correlation, s_i is usually eliminated from the computation of the total score, X .

Test items having $\bar{X}_i \leq \bar{X}$ are considered "poor" items in that $r_{s_i, X} \leq 0$. In this case, examinees that correctly answer the item tend, on average, to have the same or lower scores on Y than examinees giving an incorrect response to the item. Item difficulty, p_i , and item discrimination, $r_{s_i, X}$, are included in the "Item Analysis" section of each [Examination Report](#) prepared by University Testing Services.

Incorrect Choices

The incorrect choices or answers of a multiple-choice item are called distractors. We expect the distractors of an item to be equally attractive to the student who does not know the correct answer. As such, the proportions of examinees selecting the distractors of an item should be approximately equal. Distractors having very high or very low p values relative to the other distractors of the item are of questionable usefulness and should either be replaced or modified. University Testing Services reports the response proportions of the distractors for each item.

Item-Characteristic Curve

A very attractive way of representing the information in an item is by plotting its item-characteristic curve (ICC). ICC's play a prominent role in Item Response Theory. The ICC of an item is a plot, over all values of Y , of the probability that an examinee at point y on the underlying Y continuum correctly answers the item, i.e., $ICC_i = \Pr\{S_i = 1|Y = y\}$ for all y . Formally, the ICC for the i -th item is considered to be a normal ogive function of Y (the familiar S-shaped curve). When plotted for a sample of size N , the ICC displays the proportion of examinees at each level of Y that correctly answers the item.

The item-characteristic curve of an item reflects both its difficulty (P_i) and discrimination ($r_{s,x}$)—as well as random guessing for selected-response item types such as multiple-choice items. Items with "steep" ICC's are very useful in differentiating examinees along the Y continuum. Items having horizontal or nearly horizontal ICC's, on the other hand, provide little or no information about Y regardless of their difficulty levels. Item ICC's should always be considered when evaluating and selecting items for a test.

Item-Item Correlation

The final item characteristic we consider is the correlation coefficient for each pair of items. The correlation coefficient (phi coefficient) between dichotomous items g and h is defined as

$$\rho_{gh} = \frac{p_{gh} - p_g p_h}{\sqrt{p_g(1-p_g)} \sqrt{p_h(1-p_h)}},$$

where $p_{gh} = \Pr\{S_g = 1, S_h = 1\}$.

Our understanding of the item response process implies that each test item—certainly a "good" item—should be saturated with the underlying attribute represented by Y . As such, we expect the correlation coefficients for distinct pairs of test items to be positive and generally high. Furthermore, for fixed $Y = y$,

$$p_{gh} = \Pr\{S_g = 1, S_h = 1\} = \Pr\{S_g = 1|y\} \Pr\{S_h = 1|y\},$$

$$\rho_{gh,y} = 0,$$

i.e., the partial correlation coefficient between item g and item h for fixed Y is zero. This result leads to $\rho_{gh} = \rho_{gx} \rho_{hx}$ which relates item correlation to item discrimination (recall that total score, X , is a proxy for Y). The quantity ρ_{gh} is fundamental to the measurement of the reliability of a test.

Test Reliability

Reliability, because it reflects the precision of a measurement, is an important property of a test. In the context of testing, precision refers to the stability of an examinee's score over repeated and identical administrations of the test. From Classical Test Theory the reliability of a test is measured by the correlation coefficient between the observed scores of two parallel forms of the test, $\rho_{xx'}$ (see the document "Academic Testing: [Classical Test Theory Approach](#)"). This correlation is estimated in practice in a variety of ways.

Test-Retest Reliability

One obvious method of estimating the reliability of a test is to administer the same test on two different occasions and compute the correlation coefficient for the total test scores. This method of reliability estimation is termed test-retest reliability. There are several problems associated with the test-retest method. Carry-over effects, changes in the individual's knowledge and skills, and other effects associated with time can lead to biased estimates of test reliability.

Parallel-Forms Reliability

If parallel forms of a test are available (see the document "Academic Testing: [Classical Test Theory Approach](#)"), then the reliability of the test is estimated by the correlation coefficient for the total scores of the two test forms, $\rho_{xx'}$. This method of reliability estimation is termed parallel-forms reliability. In practice, only the largest of testing organizations have the resources to construct parallel forms of a test. Parallel-forms reliability estimates are subject to the same time-related threats associated with test-retest reliability.

Internal-Consistency Reliability

The most widely used method of reliability estimation is known as internal consistency reliability. Internal consistency reliability is attractive because it requires only one administration of the test. It is the method employed by University Testing Services in computing test reliability.

The easiest way to implement internal consistency reliability is to simply divide, in some way, the total test into two "parallel" halves denoted Y_1 and Y_2 , compute the correlation coefficient for the scores on the two halves, and then apply the Spearman-Brown formula to $\rho_{Y_1Y_2}$ (see the document "Academic Testing: [Classical Test Theory Approach](#)"). This method of reliability estimation is known as split-half reliability. Obviously, there are many ways to divide a test into two parts, but none are likely to produce the parallel measures required by the Spearman-Brown formula.

There is no reason to limit division of the total test into halves. In fact, each item can be taken as a component of the test and the relationships among the items can be used to estimate the reliability of the test. If the k items of a test are highly inter-correlated, then the reliability of the test will be high. This result is consistent with the assumption that each item is saturated with a common underlying attribute of the examinees that is measured by X . Recall that $\rho_{gh} = \rho_{gx} \rho_{hx}$. It is in this sense that the test is said to be internally consistent.

The reliability coefficient computed and reported by University Testing Services depends only upon the number of items in the test and the variances and covariances of the items. For a test of k dichotomously scored items, reliability is defined as

$$\rho_{xx'} \geq \text{KR 20} = \left(\frac{k}{k-1} \right) \left(\frac{\sigma_x^2 - \sum_{i=1}^k p_i(1-p_i)}{\sigma_x^2} \right),$$

where σ_x^2 is the population variance of the total test score and p_i is the proportion of examinees that correctly answers the i -th item. The quantity KR20 is known as the Kuder-Richardson Formula 20 (Kuder & Richardson, 1937). Another name for the formula is Coefficient α . If the test items are "essentially τ -equivalent" (a relaxed condition of parallel forms), then $\text{KR20} = \rho_{xx'}$. Otherwise, KR20 underestimates the reliability of the test. When the KR20 estimate for a test is high, we are reasonably assured that the reliability of the test is high. Low values of KR20, however, may or may not reflect low $\rho_{xx'}$ and should be interpreted with caution.

Standard Error of Measurement

Another way of expressing the precision of a total test score that is derived from Classical Test Theory is the standard error of measurement (see the document "Academic Testing: [Classical Test Theory Approach](#)"). From CTT, an examinee's observed test score (X) has two unobserved components—a true score component (T) and an error component (E). These test scores are related as $X = T + E$. Obviously, the smaller the error the greater the precision in measuring the underlying attribute. Under the assumptions of CTT, the size of the error in a measurement X is given by the standard error of measurement

$$\sigma_E = \sigma_x \sqrt{1 - \rho_{xx'}}.$$

A sample estimate of σ_E is obtained by replacing σ_x and $\rho_{xx'}$ by their sample estimates s_x and $r_{xx'}$, respectively.

We use the standard error of measurement to construct a $100(1-\alpha)\%$ confidence interval for the true score, T . Assuming normally distributed errors, a $100(1-\alpha)\%$ confidence interval for T is given by

$$(x - z_\alpha s_E \leq T \leq x + z_\alpha s_E),$$

where z_α is the upper α -point of the distribution of the standard normal variable. Before any samples are drawn, the probability that this interval contains the true score T is $1-\alpha$. As the reliability of the test increases, the confidence interval for T shortens and approaches T in the limit.

While error analysis is a powerful way of expressing the precision of a test score, s_E and the $100(1-\alpha)\%$ confidence interval for T are not included in the University Testing Services' [Examination Report](#). Everyday experience indicates that many test users do not understand either the notion of true scores or the computation and meaning of standard errors and confidence intervals.

Item Selection

The selection of items for inclusion in a test designed to measure learning outcomes is guided by the content and statistical specifications contained in the Measurement Plan for that academic course. As a general rule it is recommended that items be chosen for a test that (1) meet the content requirements of the measurement plan, (2) have the "steepest" item-characteristic curves, (3) have item difficulty values consistent with the purposes of the test, and (4) are internally consistent.

Test Score Transformation

The raw or total test score [X = the sum of the (0-1) item scores], because it depends upon the number and type of items making up a test, is generally not very informative. The raw scores of tests of different lengths, for example, are not directly comparable. It is common practice to apply some form of transformation to the raw score of a test in an effort to make the test results more meaningful, particularly to the student.

Formula Scores

A major problem associated with multiple-choice tests is that of guessing, i.e., selecting the correct response to a test item in the absence of knowledge simply by chance. Under certain assumptions the effects of guessing can be estimated and the raw score adjusted. Assuming that the examinee randomly guesses whenever he or she does not know the answer to a test question, the number of items that the examinee would answer correctly without guessing is estimated as

$$X_c = \left(\frac{k}{k-1} \right) X - \frac{N}{k-1},$$

where X_c is the raw score corrected for guessing, X is the unadjusted raw score, N is the total number of items, and k is the number of response choices for each item. When the examinee answers all the test items, X_c is a linear function of X . If examinees do not answer one or more test items, then a better formula for X_c is

$$X_c = \left(\frac{k-1}{k} \right) X + \frac{N-t}{k},$$

where t is the number of incorrect responses.

The assumption of random guessing in the absence of knowledge is a strong one and little is known of its effect upon \bar{X} when it is violated. Adjusting total test scores for guessing should be done with care.

Percentage Scores

The simplest type of raw score transformation is to divide the raw score, X , by the total number of test items, N , and multiply the result by 100. This transformed score is called the percentage score. If the items of the test are a "good" representation of the material for this part of the course, then the percentage score can be viewed as a measure of the amount of course material the student has mastered. Obviously, interpretation of a percentage score is highly dependent upon the nature of the items making up the test, i.e. item content, item difficulty, and item discrimination. On the other hand, properties of the students taking the test do not enter into the interpretation of the percentage score.

Percentile Scores

Percentiles are transformed scores that are based upon the test performance of a specific group of individuals often called the norm group. The percentile score or rank is defined as the percentage of individuals in the norm group having a score equal to or less than a particular value. There are several ways of computing percentile scores. A typical method is to take one-half the number of individuals obtaining the specified score, add that number to the number of individuals obtaining a lower score, and divide the sum by the total number of individuals. However it is computed the percentile score is interpreted in reference to the norm group. Change the norm group and the interpretation of the percentile score changes.

While it is relatively easy to compute and interpret, the percentile score has its limitations. For one, the distribution of percentiles is rectangular and not amenable to many common statistical analyses. For another, differences among percentile scores can be misleading particularly when a large number of individuals in the norm group have the same or similar scores (as is the case with normally distributed raw scores).

Standardized Scores

Another very popular type of norm-referenced score is the standardized score. For a "norm" population having mean μ_x and variance σ_x^2 , the standardized score of an examinee with score X_i is

$$Y_i = \sigma^* \left(\frac{X_i - \mu_x}{\sigma_x} \right) + \mu^*,$$

where, μ^* and σ^* are arbitrary. For example, the College Entrance Examination Board's SAT Verbal and Quantitative scores each have $\mu^* = 500$ and $\sigma^* = 100$. Since it is a linear transformation of X , Y has the same distribution as X . If that distribution is approximately normal, then the standardized score can be interpreted in reference to the standard normal deviate, i.e., interpreted as deviations about the mean. This may not be a useful interpretation for those unskilled in statistics.

Normalized Scores

As the name implies, the normalized score involves a transformation of the raw score such that the normalized score follows the normal distribution regardless of the distribution of X . While tedious to compute, the transformation is straightforward--first transforming X to percentiles and then finding the unit normal deviate corresponding to the each percentile. In most cases, the resulting normalized score is standardized to have mean μ^* and standard deviation σ^* . The two most common normalized scores are the T Score for which $\mu^* = 50$ and $\sigma^* = 10$, and the Stanine having $\mu^* = 5$ and σ^* approximately equal to 2.

Comment

In most applied instructional settings, e.g., academic courses at the high school or college level, it is not common to have an appropriately specified norm or reference group on which to base test score transformations. For this reason, University Testing Services only reports the raw score and the corresponding percentage score in the [Examination Report](#) for each examination in a course.

References

- Gronlund, N. E. (1982). Constructing achievement tests. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
- Haladyna, T. M. (1999) Developing and validating multiple-choice test items. 2nd ed. Mahwah, NJ: Erlbaum.
- Kuder, G. F. & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160.

<http://webusers.xula.edu/jsevenai/objective/guidelines.html>



Item Writing Guidelines

A point to keep in mind is that you can compensate for ambiguities and misphrasing in grading numerical problems and essays. For multiple choice items, on the other hand, you must apply grammar and logic rigorously if your questions are to be useful. It takes considerable time and thought to construct a good multiple choice item. Writing well-phrased stems with plausible foils is hardly ever easy.

The guidelines presented here have been gleaned from a variety of sources (6-10). The most important source has unquestionably been practical experience.

In reading these guidelines over, you might note that they can be distilled into two statements: (1) Don't make your questions too hard (by introducing irrelevant complications). (2) Don't make your questions too easy (by failing to tap higher-level cognitive skills).

General Considerations

A. Be Clear and Concise.

You can't overstate the importance of clarity. Think of this as a rule, not a guideline. **BE CLEAR.**

Without sacrificing clarity, be as concise as possible. Keep a sense of focus. Remember that your purpose is to measure your students' knowledge, reasoning, and ability. Verbal gamesmanship is pointless. The idea is to discriminate levels of understanding, not to trap the unwary.

Simply stated: It is not sufficient to write a question that can be understood. To paraphrase the advice of Gen. Douglas MacArthur to the Corps of Cadets: Write questions that cannot be misunderstood, not merely questions that can be understood.

B. Use the Active Voice.

Everybody processes information much more easily in the active voice. There is a host of research to support this contention (11-19); two particularly fine sources are Olsen (17) and Tichy (18).

C. Watch Difficulty Levels

The ideal question will be answered correctly by 60-65% of the tested population. This level of difficulty maximizes discrimination on exams.

In the sciences at least, it can be an adventure to write items that are this easy. Instructors tend to overestimate student abilities badly. Many item writers use their own capabilities as a yardstick, and forget the years of practice that went into honing their skills. Whatever else you try, don't try to be too clever!

D. Tap Higher Level Cognitive Domains

Rote memorization of facts, laws, and definitions has its place in the overall scheme. However, at least 90% of the test should be devoted to higher levels of cognition.

E. Get everything peer-reviewed.

This technique is unbeatable in trouble-shooting and improving composed items. There is considerable merit to constructive criticism. You wouldn't submit a paper for publication without letting a colleague take a look at it, would you?

Also, remember that writing is a difficult task for EVERYBODY, with the possible exception of Isaac Asimov. Preparing good multiple choice items is a scholarly activity that demands time, clarity of thought, and precision in expression.

Remember that students read test items more carefully than they read anything else. All flaws and imperfections will be exposed.

Reproduced here, without comment, are some "Rules of English" that were published anonymously in the Chronicle of Higher Education (May 19, 1982).

1. Don't use no double negatives.
2. Make each pronoun agree with their antecedent.
3. Join clauses good, like a conjunction should.
4. When dangling watch them participles.
5. About them sentence fragments.
6. Verbs has to agree with their subject.
7. Just between you and I, case is important, too.
8. Don't write run-on sentences they are hard to read.
9. Don't use commas, which are not necessary.
10. Try to not ever split infinitives.
11. Its important to use your apostrophe's correctly.
12. Proof read your writing to see if any words out.

F. Be Clear And Concise (reminder).

Writing Stems**A. Pose a Question or an Incomplete Thought.**

Items should ask direct and complete questions. The task must be clear; students must not be forced to read all the responses in order to know what question is being posed. You know you're on track if the student can read the stem, formulate the answer, then pattern-match this answer to the key.

Example

Poor: A mammal:

A. Duck B. Lizard C. Cat * D. Trout

Improved: Which of the following is a mammal?

B. Using "Which of the following . . ." and related stems.

Items that begin "Which of the following . . ." are also acceptable when the task is well- defined. However, don't use "Which of the following . . ." when the answer is unique.

Example

Poor: Which of the following is true?

Improved: Which of the following statements about the structure of a cell is true?

Poor: Which of the following is the symbol for tin?

Improved: What is the chemical symbol for tin?

The wording "Which of the following" is considered somewhat old-fashioned in some circles. The concept remains tremendously useful, however. Possible alternative wordings are "Which of these" and "Which of the (objects, statements, etc.) shown".

C. Stems Containing Blanks

Using blanks is also acceptable. The blanks should appear at the end of the stem, not at the beginning.

Example

Poor: The molar mass of X is _____ the molar mass of Y if molecule X diffuses twice as fast as molecule Y.

Improved: Molecule X diffuses twice as fast as molecule Y when the molar mass of X is _____ the molar mass of Y.

- A. one-fourth *
- B. one-half
- C. twice
- D. four times

D. Focus on significant or important concepts.

Avoid trivia and unimportant details. This states the obvious, perhaps, but this is a guideline that is violated all too often. Which student is more likely to succeed, one with a vast memorized knowledge of minute detail or one with the ability to use classroom knowledge in new situations?

Example

Poor: What is a Lucas test?

Improved: Which of the following compounds reacts rapidly with $\text{ZnCl}_2/\text{HCl}(\text{aq})$?

E. Minimize the use of negatives.

Items addressed to exceptions are acceptable. "Which of the following is NOT . . ." is an appropriate stem, but it must not be overused. When you do use this type of stem, you must strongly emphasize the **NEGATIVE**. All caps and either underlining or boldface is the recommended style. Of course double negatives (negative stems and negative foils in the same item) must be avoided completely.

Example

Poor: None of the following modifications increases the yield of Reaction 2 except:

Improved: Which of the following modifications increases the yield of Reaction 2?

F. Don't be wordy.

Don't make your stems too wordy or unnecessarily complicated. Item stems that exceed 50 words are unacceptable.

G. Do not key the answer in the stem.

This guideline is all too easy to violate, and there are many ways to get trapped. By way of illustration, in the question, "The structure shown is a: A. Ketone, B. Aldehyde, C. Ester, D. Alcohol," only foil A is grammatically correct and must be the keyed response. (Notice that this item can be answered without looking at the structure!).

Another thing to avoid is the use of words that can key the response. For example, the key to the question "Which of the following is a chemical property of sulfur?" should not contain the word *reaction*. The words *chemical* and *reaction* are too closely related.

Example

Which of the following most likely results from evolution?

Poor foil: A. Development of antibiotic resistance in bacteria.

Note that evolution and development are closely aligned.

Improved foil: A. Antibiotic resistance in bacteria.

Foils

A. They must flow smoothly from the stem

All foils must be in the same format phrases, names, numbers, etc. Responses must be grammatically correct and logically consistent with the stem.

Example

Poor: Which of the following compounds is a stronger base than NaOH?

I. NaH II. NaF III. Na₂SO₄

A. I only * B. II only C. I and III only D. I, II, and III

Note: The word *is* in the stem tells the students that only one of the compounds is a stronger base than NaOH. Testwise students will consider only foils A and B.

B. Use plausible distractors

Your distractors must be plausible in terms of the question asked. True statements generally make very effective distractors.

Here's a particularly useful method for obtaining good distractors. Clearly state to yourself how a poorly-prepared student can arrive at each distractor.

Example

Poor: A diet low in iodine can result in an enlarged thyroid because:

- A. iodine is toxic to the thyroid.
- B. the thyroid removes iodine from the blood.
- C. the thyroid hormone requires iodine to function. *
- D. iodine is produced in the thyroid.

Note: A student who knows no biology can answer this one. The situations in foils A, B, and D shouldn't pose a problem in a low-iodine environment. Foils must have surface plausibility if the item is to be effective.

C. Never use "All of the above."

If a student recognizes two of the alternatives as correct, then logic dictates that "All of the above" is the answer. Also, students who recognize one of the alternatives as incorrect will know that "All of the above" can't be the correct answer. Furthermore, if students chooses one of the correct answers from an "All of the above" set, they can argue with some justification that they should get credit (see section E below).

D. Never use "None of the above."

When "None of the above" is used, the student must formulate a response to compare with the foils. If the instructor's key has any problems at all, "None of the above" becomes an arguably correct response. Furthermore, it is probably true that instructors most often use "None of the above" when they can't think of a plausible last foil to meet a predetermined number. There are a few situations where "None of the above" can be used effectively, but it is better to avoid the problem entirely.

E. There must be only one correct answer.

The item must have one, and only one, acceptable response. No distractor can be close enough to the correct answer to be arguable. The foils **MUST** be mutually exclusive.

Example

Poor: Which of the following bases is (are) found in RNA?

A. Adenine B. Guanine C. Uracil D. All of these *

A test-wise student will be led directly to the key by recognizing only two RNA bases. Also note that foils A, B, and C are arguable since these bases are, indeed, found in RNA. Foils must be mutually exclusive.

Improved: Which of the following bases is (are) found in RNA?

I. Uracil II. Guanine III. Adenine

A. I only B. I and II only C. II and III only D. I, II, and III

F. Use homogeneous foils

If one of your foils is markedly different from the others, you will distract the students from the content-area problem. It is irrelevant whether this difference is in content (e.g. apple, pear, peach, cat) or length. The difficulties are much more serious if the distinctive answer is consistently the correct response.

If foils are of different lengths, then use two long and two short foils. (It is also a little better to put the two long foils adjacent to each other and the two short ones adjacent to each other.) If you use pairs of opposites (see section K below), then use two pairs.

Examples

Poor: A likely source of nourishment for humans in regions of the world where food is scarce is:

A. beef.
B. plants such as rice, corn, beans, and wheat. *
C. pork.
D. algae.

Improved foils, using pairs of opposites and grouping the pairs:

A. beef B. pork. C. grain.* D. algae.

Poor: One problem associated with the destruction of the ozone layer is the killing of phytoplankton at the ocean's surface. This would likely result in all of the following EXCEPT:

A. lowering of fish populations.
B. lowering marine mammal numbers.

- C. increasing the dissolved oceanic O₂ levels. *
- D. decreasing oceanic invertebrate populations.

Note that there are three "getting smaller" and one "getting larger", three organisms and one molecule. The distinctive foil stands up, waves a flag, and screams "PICK ME! PICK ME!!" to the test-wise student.

G. Avoid overlapping responses

You should avoid beginning or ending a set of foils with identical words or phrases. If all of your foils begin with the same word or phrase, you should put that portion of the foil in the stem. There is no advantage in requiring the students to read the same phrase four times it's merely time-consuming.

Example

Poor: What is the major effect on a cell if it is placed in a solution that inhibits its ribosomes?

- A. The cell would not be able to produce ATP.
- B. The cell would not be able move about.
- C. The cell could not synthesize proteins.
- D. The cell could not carry out photosynthesis.

Improved: A cell placed in a solution that inhibits ribosome activity is unable to:

- A. produce ATP.
- B. move in solution.
- C. synthesize proteins.
- D. conduct photosynthesis.

H. Place foils in numerical or chronological order

When the foils are numbers or dates, these should be in order, high to low or low to high. You should not scramble numbers and force the students to search the alternatives to find the key. Exception: Instructors are sometimes faced with large numbers of students in relatively small rooms. They sometimes produce multiple test versions by scrambling the order of both the questions and the foils for security reasons. This is not ideal, but is allowable.

I. Distribute the key statistically

When preparing a test, you must have the keyed response statistically distributed over the A's, B's, C's, and D's. It is much easier to do this if your test bank has a statistical distribution of responses. You can save yourself a lot of grief by keeping track of the key distribution as you go along. Many item writers underuse foil D; A is also a problem sometimes.

One particular difficulty here is with numerical foils. Take some care to make certain that both the largest and the smallest numbers share key-work evenly with the rest of the numbers. A major league sinner is giving you this advice.

J. Number of Foils to Use

The number of foils is relatively unimportant. Three-, four-, and five-foil questions are all used on some standard examinations.

Theoretically, you achieve the maximum discrimination when the average score on an exam is exactly halfway between a perfect score and a "pure guess" result. This is an average of 60% for a five-foil test, 62.5% for a four-foil test, and 66.7% for a three-foil test. The difference will not be statistically significant on any test we can construct.

K. What Makes Effective Foils?

True statements make very effective foils in all situations.

Another good technique is to use "pairs of opposites". This is particularly effective when TWO pair of opposites are used. For example : Increasing and decreasing temperature paired with increasing and decreasing pressure. The foil sets do not have to be "true" opposites two mammals and two reptiles, two planets and two stars, two bodies of salt water and two of fresh water, etc., work very well.

Item writing checklist

1. Is the item clear and concise?
2. Did you use the active voice?
3. Did you avoid "ould" words?
4. Is the difficulty level acceptable?
5. Does the stem pose a question or an incomplete thought?
6. If you used blanks, are they at the end of the stem?
7. Does the stem focus on a significant or important aspect?
8. Did you emphasize the NEGATIVES?
9. Have you avoided keying the answer in the stem?
10. Are the distractors plausible?
11. Is there only one arguable correct response?
12. Are the foils homogeneous?
13. Did you avoid overlapping foils?
14. Are numerical foils in either ascending or descending order?

Go to the | [Objective Testing](#) | [John P. Sevenair](#) | [Xavier University](#) | [Home Page](#)

<http://www.oir.uiuc.edu/dme/exams/ITQ.html>

Improving Your Test Questions

Table of Contents

[Choosing between Objective and Subjective Test Items](#)

[Suggestions for Using and Writing Test Items](#)

[Multiple Choice](#)

[True-False](#)

[Matching](#)

[Completion](#)

[Essay](#)

[Problem Solving](#)

[Performance](#)

[Two Methods for Assessing Test Item Quality](#)

[Assistance Offered by The Office of Instructional Resources \(OIR\)](#)

[References for Further Reading](#)

I. CHOOSING BETWEEN OBJECTIVE AND SUBJECTIVE TEST ITEMS

II.

There are two general categories of test items: (1) objective items which require students to select the correct response from several alternatives or to supply a word or short phrase to answer a question or complete a statement; and (2) subjective or essay items which permit the student to organize and present an original answer. Objective items include multiple-choice, true-false, matching and completion, while subjective items include short-answer essay, extended-response essay, problem solving and performance test items. For some instructional purposes one or the other item types may prove more efficient and appropriate. To begin our discussion of the relative merits of each type of test item, test your knowledge of these two item types by answering the following questions.

Test Item Quiz

	(circle the correct answer)		
1. Essay exams are easier to construct than are objective exams.	T	F	?
2. Essay exams require more thorough student preparation and study time than objective exams.	T	F	?
3. Essay exams require writing skills where objective exams do not.	T	F	?
4. Essay exams teach a person how to write.	T	F	?
5. Essay exams are more subjective in nature than are objective exams.	T	F	?
6. Objective exams encourage guessing more so than essay exams.	T	F	?
7. Essay exams limit the extent of content covered.	T	F	?
8. Essay and objective exams can be used to measure the same content or ability.	T	F	?

9. Essay and objective exams are both good ways to evaluate a student's level of knowledge. T F ?

Quiz Answers

1. TRUE Essay items are generally easier and less time consuming to construct than are most objective test items. Technically correct and content appropriate multiple-choice and true-false test items require an extensive amount of time to write and revise. For example, a professional item writer produces only 9-10 good multiple-choice items in a day's time.
2. ? According to research findings it is still undetermined whether or not essay tests require or facilitate more thorough (or even different) student study preparation.
3. TRUE Writing skills do affect a student's ability to communicate the correct "factual" information through an essay response. Consequently, students with good writing skills have an advantage over students who have difficulty expressing themselves through writing.
4. FALSE Essays do not teach a student how to write but they can emphasize the importance of being able to communicate through writing. constant use of essay tests may encourage the knowledgeable but poor writing student to improve his/her writing ability in order to improve performance.
5. TRUE Essays are more subjective in nature due to their susceptibility to scoring influences. Different readers can rate identical responses differently, the same reader can rate the same paper differently over time, the handwriting, neatness or punctuation can unintentionally affect a paper's grade and the lack of anonymity can affect the grading process. While impossible to eliminate, scoring influences or biases can be minimized through procedures discussed later in this booklet.
6. ? Both item types encourage some form of guessing. Multiple-choice, true-false and matching items can be correctly answered through blind guessing, yet essay items can be responded to satisfactorily through well written bluffing.
7. TRUE Due to the extent of time required by the student to respond to an essay question, only a few essay questions can be included on a classroom exam. Consequently, a larger number of objective items can be tested in the same amount of time, thus enabling the test to cover more content.
8. TRUE Both item types can measure similar content or learning objectives. Research has shown that students respond almost identically to essay and objective test items covering the same content. Studies¹ by Sax & Collet (1968) and Paterson (1926) conducted forty-two years apart reached the same conclusion:

"...there seems to be no escape from the conclusions that the two types of exams are measuring identical things." (Paterson, p. 246)

This conclusion should not be surprising; after all, a well written essay item requires that the student (1) have a store of knowledge, (2) be able to relate facts and principles, and (3) be able to organize such information into a coherent and logical written expression, whereas an objective test item requires that the student (1) have a store of knowledge, (2) be able to relate facts and principles, and (3) be able to organize such information into a

coherent and logical choice among several alternatives.

9. TRUE Both objective and essay test items are good devices for measuring student achievement. However, as seen in the previous quiz answers, there are particular measurement situations where one item type is more appropriate than the other. Following is a set of recommendations for using either objective or essay test items: (Adapted from Robert L. Ebel, *Essentials of Educational Measurement*, 1972, p. 144).

¹Gilbert Sax and LeVerne S. Collet, "An Empirical Comparison of the Effects of Recall and Multiple-Choice Tests on Student Achievement," *Journal of Educational Measurement*, vol. 5 (1968), 169-73.
Donald G. Paterson, "Do New and Old Type Examinations Measure Different Mental Functions?" *School and Society*, vol. 24. (August 21, 1926), 246-48.

WHEN TO USE ESSAY OR OBJECTIVE TESTS

Essay tests are especially appropriate when:

- the group to be tested is small and the test is not to be reused.
- you wish to encourage and reward the development of student skill in writing.
- you are more interested in exploring the student's attitudes than in measuring his/her achievement.
- you are more confident of your ability as a critical and fair reader than as an imaginative writer of good objective test items.

Objective tests are especially appropriate when:

- the group to be tested is large and the test may be reused.
- highly reliable test scores must be obtained as efficiently as possible.
- impartiality of evaluation, absolute fairness, and freedom from possible test scoring influences (e.g., fatigue, lack of anonymity) are essential.
- you are more confident of your ability to express objective test items clearly than of your ability to judge essay test answers correctly.
- there is more pressure for speedy reporting of scores than for speedy test preparation.

Either essay or objective tests can be used to:

- measure almost any important educational achievement a written test can measure.
- test understanding and ability to apply principles.
- test ability to think critically.
- test ability to solve problems.
- test ability to select relevant facts and principles and to integrate them toward the solution of complex problems.

In addition to the preceding suggestions, it is important to realize that certain item types are **better suited** than others for measuring particular learning objectives. For example, learning objectives requiring the student **to demonstrate** or **to show**, may be better measured by performance test items, whereas objectives requiring the student **to explain** or **to describe** may be better measured by essay test items. The matching of learning objective expectations with

certain item types can help you select an appropriate kind of test item for your classroom exam as well as provide a higher degree of test validity (i.e., testing what is supposed to be tested). To further illustrate, several sample learning objectives and appropriate test items are provided on the following page.

Learning Objectives	Most Suitable Test Item
The student will be able to categorize and name the parts of the human skeletal system.	Objective Test Item (M-C, T-F, Matching)
The student will be able to critique and appraise another student's English composition on the basis of its organization.	Essay Test Item (Extended-Response)
The student will demonstrate safe laboratory skills.	Performance Test Item
The student will be able to cite four examples of satire that Twain uses in <i>Huckleberry Finn</i> .	Essay Test Item (Short-Answer)

After you have decided to use either an objective, essay or both objective and essay exam, the next step is to select the kind(s) of objective or essay item that you wish to include on the exam. To help you make such a choice, the different kinds of objective and essay items are presented in the following section of this booklet. The various kinds of items are briefly described and compared to one another in terms of their advantages and limitations for use. Also presented is a set of general suggestions for the construction of each item variation.

II. MULTIPLE-CHOICE TEST ITEMS

The multiple-choice item consists of two parts: (a) the stem, which identifies the question or problem and (b) the response alternatives. Students are asked to select the one alternative that best completes the statement or answers the question. For example,

Sample multiple-Choice Item

(a) *Item Stem: Which of the following is a chemical change?*

- (b) *Response Alternatives:*
- a. Evaporation of alcohol*
 - b. Freezing of water*
 - *c. Burning of oil*
 - d. Melting of wax*

*correct response

Advantages in Using Multiple-Choice Items

Multiple-choice items can provide ...

- versatility in measuring all levels of cognitive ability.
- highly reliable test scores.
- scoring efficiency and accuracy.

- objective measurement of student achievement or ability.
- a wide sampling of content or objectives.
- a reduced guessing factor when compared to true-false items.
- different response alternatives which can provide diagnostic feedback.

Limitations in Using Multiple-Choice Items

Multiple-choice items ...

- are difficult and time consuming to construct.
- lead an instructor to favor simple recall of facts.
- place a high degree of dependence on the student's reading ability and instructor's writing ability.

SUGGESTIONS FOR WRITING MULTIPLE-CHOICE TEST ITEMS

The Stem

1. When possible, state the stem as a direct question rather than as an incomplete statement.
Undesirable: *Alloys are ordinarily produced by ...*
Desirable: *How are alloys ordinarily produced?*
2. Present a definite, explicit and singular question or problem in the stem.
Undesirable: *Psychology ...*
Desirable: *The science of mind and behavior is called ...*
3. Eliminate excessive verbiage or irrelevant information from the stem.
Undesirable: *While ironing her formal, Jane burned her hand accidentally on the hot iron. This was due to a transfer of heat be ...*
Desirable: *Which of the following ways of heat transfer explains why Jane's hand was burned after she touched a hot iron?*
4. Include in the stem any word(s) that might otherwise be repeated in each alternative.
Undesirable: *In national elections in the United States the President is officially*
a. chosen by the people.
b. chosen by members of Congress.
c. chosen by the House of Representatives.
**d. chosen by the Electoral College.*
Desirable: *In national elections in the United States the President is officially chosen by*

- a. the people.*
- b. members of Congress.*
- c. the House of Representatives.*
- *d. the Electoral college.*

5. Use negatively stated stems sparingly. When used, underline and/or capitalize the negative word.

Undesirable: *Which of the following is not cited as an accomplishment of the Kennedy administration?*

Desirable: *Which of the following is NOT cited as an accomplishment of the Kennedy administration? Item Alternatives*

6. Make all alternatives plausible and attractive to the less knowledgeable or skillful student.

What process is most nearly the opposite of photosynthesis?

Undesirable Desirable

a. Digestion

b. Relaxation

**c. Respiration*

d. Exertion

a. Digestion

b. Assimilation

**c. Respiration*

d. Catabolism

7. Make the alternatives grammatically parallel with each other, and consistent with the stem.

Undesirable: *What would do most to advance the application of atomic discoveries to medicine?*

**a. Standardized techniques for treatment of patients.*

b. Train the average doctor to apply radioactive treatments.

c. Remove the restriction on the use of radioactive substances.

d. Establishing hospitals staffed by highly trained radioactive therapy specialists.

Desirable: *What would do most to advance the application of atomic discoveries to medicine?*

**a. Development of standardized techniques for treatment of patients.*

b. Training of the average doctor in application of radioactive treatments.

c. Removal of restriction on the use of radioactive substances.

d. Addition of trained radioactive therapy specialists to hospital staffs.

8. Make the alternatives mutually exclusive.

Undesirable: *The daily minimum required amount of milk that a 10 year old child should drink is*

- a. 1-2 glasses.
- *b. 2-3 glasses.
- *c. 3-4 glasses.
- d. at least 4 glasses.

Desirable: *What is the daily minimum required amount of milk a 10 year old child should drink?*

- a. 1 glass.
- b. 2 glasses.
- *c. 3 glasses.
- d. 4 glasses.

9. When possible, present alternatives in some logical order (e.g., chronological, most to least, alphabetical).

At 7 a.m. two trucks leave a diner and travel north. One truck averages 42 miles per hour and the other truck averages 38 miles per hour. At what time will they be 24 miles apart?

Undesirable Desirable

- | | |
|------------|------------|
| a. 6 p.m. | a. 1 a.m. |
| b. 9 p.m. | b. 6 a.m. |
| c. 1 a.m. | c. 9 a.m. |
| *d. 1 p.m. | *d. 1 p.m. |
| e. 6 a.m. | e. 6 p.m. |

10. Be sure there is only one correct or best response to the item.

Undesirable: *The two most desired characteristics in a classroom test are validity and*

- a. precision.
- *b. reliability.
- c. objectivity.
- *d. consistency.

Desirable: *The two most desired characteristics in a classroom test are validity and*

- a. precision.
- *b. reliability.
- c. objectivity.
- d. standardization.

11. Make alternatives approximately equal in length.

Undesirable: *The most general cause of low individual incomes in the United States is*

- *a. lack of valuable productive services to sell.
- b. unwillingness to work.
- c. automation.
- d. inflation.

d. inflation.

Desirable: *What is the most general cause of low individual incomes in the United States?*

- *a. A lack of valuable productive services to sell.*
- b. The population's overall unwillingness to work.*
- c. The nation's increased reliance on automation.*
- d. an increasing national level of inflation.*

12. Avoid irrelevant clues such as grammatical structure, well known verbal associations or connections between stem and answer.

Undesirable: *A chain of islands is called an:*

(grammatical
clue)

- *a. archipelago.*
- b. peninsula.*
- c. continent.*
- d. isthmus.*

Undesirable: *The reliability of a test can be estimated by a coefficient of:*

(verbal
association
clue)

- a. measurement.*
- *b. correlation.*
- c. testing.*
- d. error.*

Undesirable: *The **height** to which a water dam is built depends on*

(connection
between stem
and answer clue)

- a. the length of the reservoir behind the dam.*
- b. the volume of water behind the dam.*
- *c. the **height** of water behind the dam.*
- d. the strength of the reinforcing wall.*

13. Use at least four alternatives for each item to lower the probability of getting the item correct by guessing.

14. Randomly distribute the correct response among the alternative positions throughout the test having approximately the same proportion of alternatives a, b, c, d and e as the correct response.

15. Use the alternatives "none of the above" and "all of the above" sparingly. When used, such alternatives should occasionally be used as the correct response.

TRUE-FALSE TEST ITEMS

A true-false item can be written in one of three forms: simple, complex, or compound. Answers can consist of only two choices (simple), more than two choices (complex), or two choices plus a conditional completion response (compound). An example of each type of true-false item follows:

Sample True-False Item: Simple

The acquisition of morality is a developmental process. *True* *False*

Sample True-False Item: Complex

The acquisition of morality is a developmental process. *True* *False* *Opinion*

Sample True-False Item: Compound

The acquisition of morality is a developmental process.

If this statement is false, what makes it false? *True* *False*

Advantages in Using True-False Items

True-False items can provide ...

- the widest sampling of content or objectives per unit of testing time.
- scoring efficiency and accuracy.
- versatility in measuring all levels of cognitive ability.
- highly reliable test scores.
- an objective measurement of student achievement or ability.

Limitations in Using True-False Items

True-false items ...

- incorporate an extremely high guessing factor. For simple true-false items, each student has a 50/50 chance of correctly answering the item without any knowledge of the item's content.
- can often lead an instructor to write ambiguous statements due to the difficulty of writing statements which are unequivocally true or false.
- do not discriminate between students of varying ability as well as other item types.
- can often include more irrelevant clues than do other item types.
- can often lead an instructor to favor testing of trivial knowledge.

SUGGESTIONS FOR WRITING TRUE-FALSE TEST ITEMS

1. Base true-false items upon statements that are absolutely true or false, without qualifications or exceptions.

- Undesirable: *Nearsightedness is hereditary in origin.*
- Desirable: *Geneticists and eye specialists believe that the predisposition to nearsightedness is hereditary.*
2. Express the item statement as simply and as clearly as possible.
- Undesirable: *When you see a highway with a marker that reads, "Interstate 80" you know that the construction and upkeep of that road is built and maintained by the state and federal government.*
- Desirable: *The construction and maintenance of interstate highways is provided by both state and federal governments.*
3. Express a single idea in each test item.
- Undesirable: *Water will boil at a higher temperature if the atmospheric pressure on its surface is increased and more heat is applied to the container.*
- Desirable: *Water will boil at a higher temperature if the atmospheric pressure on its surface is increased.*
and/or
Water will boil at a higher temperature if more heat is applied to the container.
4. Include enough background information and qualifications so that the ability to respond correctly to the item does not depend on some special, uncommon knowledge.
- Undesirable: *The second principle of education is that the individual gathers knowledge.*
- Desirable: *According to John Dewey, the second principle of education is that the individual gathers knowledge.*
5. Avoid lifting statements from the text, lecture or other materials so that memory alone will not permit a correct answer.
- Undesirable: *For every action there is an opposite and equal reaction.*
- Desirable: *If you were to stand in a canoe and throw a life jacket forward to another canoe, chances are your canoe would jerk backward.*
6. Avoid using negatively stated item statements.
- Undesirable: *The Supreme Court is not composed of nine justices.*
- Desirable: *The Supreme is composed of nine justices.*
7. Avoid the use of unfamiliar vocabulary.
- Undesirable: *According to some politicians, the raison d'etre for capital punishment is retribution.*
- Desirable: *According to some politicians, justification for the existence of capital punishment is retribution.*
8. Avoid the use of specific determiners which would permit a test-wise but unprepared examinee to respond correctly. Specific determiners refer to sweeping terms like "all," "always," "none," "never," "impossible," "inevitable," etc. Statements including such terms are likely to be false. On the other hand, statements using qualifying determiners such as

"usually," "sometimes," "often," etc., are likely to be true. When statements do require the use of specific determiners, make sure they appear in both true and false items.

Undesirable: *All sessions of Congress are called by the President. (F)*

*The Supreme Court is **frequently** required to rule on the constitutionality of a law. (T)*

*An objective test is **generally** easier to score than an essay test. (T)*

Desirable: *(When specific determiners are used reverse the expected outcomes.)*

*The sum of the angles of a triangle is **always** 180°. (T)*

*Each molecule of a given compound is chemically the same as **every** other molecule of that compound. (T)*

*The galvanometer is the instrument **usually** used for the metering of electrical energy used in a home. (F)*

9. False items tend to discriminate more highly than true items. Therefore, use more false items than true items (but no more than 15% additional false items).

MATCHING TEST ITEMS

In general, matching items consist of a column of stimuli presented on the left side of the exam page and a column of responses placed on the right side of the page. Students are required to match the response associated with a given stimulus. For example,

Sample Matching Test Item

Directions: On the line to the left of each factual statement, write the letter of the principle which best explains the statement's occurrence. Each principle may be used more than once.

Factual Statements

1. Fossils of primates first appear in the Cenozoic rock strata, while trilobite remains are found in the Proterozoic rocks.
2. The Arctic and Antarctic regions are sparsely populated.
3. Plants have no nervous system.
4. Large coal beds exist in Alaska.

Principles

- a. There have been profound changes in the climate on earth.
- b. Coordination and integration of action is generally slower in plants than in
- c. There is an increasing complexity of structure and functions from lower to
- d. All life comes from life and produces its own kind of living organisms.

- d. All life comes from life and produces its own kind of living organisms.*
e. Light is a limiting factor to life.

Advantages in Using Matching Items

Matching items

- require short periods of reading and response time, allowing you to cover more content.
- provide objective measurement of student achievement or ability.
- provide highly reliable test scores.
- provide scoring efficiency and accuracy.

Limitations in Using Matching Items

Matching items

- have difficulty measuring learning objectives requiring more than simple recall of information.
- are difficult to construct due to the problem of selecting a common set of stimuli and responses.

SUGGESTIONS FOR WRITING MATCHING TEST ITEMS

1. Include directions which clearly state the basis for matching the stimuli with the responses. Explain whether or not a response can be used more than once and indicate where to write the answer.

Undesirable: *Directions: Match the following.*

Desirable: *Directions: On the line to the left of each identifying location and characteristics in Column I, write the letter of the country in Column II that is best defined. Each country in Column II may be used more than once.*

2. Use only homogeneous material in matching items.

Undesirable: *Directions: Match the following.*

- | | |
|--|---------------------|
| 1. ____ Water | A. NaCl |
| 2. ____ Discovered Radium | B. Fermi |
| 3. ____ Salt | C. NH ₃ |
| 4. ____ Year of the 1st Nuclear Fission by Man | D. H ₂ O |
| 5. ____ Ammonia | E. 1942 |
| | F. Curie |

Desirable: *Directions: On the line to the left of each compound in Column I, write the letter of the compound's formula presented in Column II. Use each formula only once.*

Column I	Column II
1. ____ Water	A. H_2SO_4
2. ____ Salt	B. HCl
3. ____ Ammonia	C. $NaCl$
4. ____ Sulfuric Acid	D. H_2O
	$_2HCl$

3. Arrange the list of responses in some systematic order if possible (e.g., chronological, alphabetical).

Directions: On the line to the left of each definition in Column I, write the letter of the defense mechanism in Column II that is described. Use each defense mechanism only once.

	Column I	Undesirable Column II	Desirable
____	1. Hunting for reasons to support one's beliefs.	a. Rationalization	a. Denial of reality
____	2. Accepting the values and norms of others as one's own even when they are contrary to previously held values.	b. Identification	b. Identification
____	3. Attributing to others one's own unacceptable impulses, thoughts and desires.	c. Projection	c. Introjection
____	4. Ignoring disagreeable situations, topics, sights.	d. Introjection	d. Projection
		e. Denial of Reality	e. Rationalization

4. Avoid grammatical or other clues to the correct response.

Undesirable: *Directions: Match the following in order to complete the sentences on the left.*

- | | |
|--|---|
| ___ 1. Igneous rocks are formed | A. a hardness of 7. |
| ___ 2. The formation of coal
requires | B. with crystalline rock. |
| ___ 3. A geode is filled | C. a metamorphic rock. |
| ___ 4. Feldspar is classified as | D. heat and pressure. |
| | E. through the solid-ification of molten
lava. |

Desirable: *Avoid sentence completion due to grammatical clues.*

5. Keep matching items brief, limiting the list of stimuli to under 10. 6. Include more responses than stimuli to help prevent answering through the process of elimination. 7. When possible, reduce the amount of reading time by including only short phrases or single words in the response list.

COMPLETION TEST ITEMS

The completion item requires the student to answer a question or to finish an incomplete statement by filling in a blank with the correct word or phrase. For example,

Sample Completion Item

According to Freud, personality is made up of three major systems, the _____, the _____ and the _____.

Advantages in Using Completion Items

Completion items

- can provide a wide sampling of content.
- can efficiently measure lower levels of cognitive ability.
- can minimize guessing as compared to multiple-choice or true-false items.

- can usually provide an objective measure of student achievement or ability.

Limitations in Using Completion Items

Completion items

- are difficult to construct so that the desired response is clearly indicated.
- have difficulty measuring learning objectives requiring more than simple recall of information.
- can often include more irrelevant clues than do other item types.
- are more time consuming to score when compared to multiple-choice or true-false items.
- are more difficult to score since more than one answer may have to be considered correct if the item was not properly prepared.

SUGGESTIONS FOR WRITING COMPLETION TEST ITEMS

1. Omit only significant words from the statement.

Undesirable: *Every atom has a central (core) called a nucleus.*

Desirable: *Every atom has a central core called a(n) (nucleus).*

2. Do not omit so many words from the statement that the intended meaning is lost.

Undesirable: *The _____ were to Egypt as the _____ were to Persia and as _____ were to the early tribes of Israel.*

Desirable: *The Pharaohs were to Egypt as the _____ were to Persia and as _____ were to the early tribes of Israel.*

3. Avoid grammatical or other clues to the correct response.

Undesirable: *Most of the United States' libraries are organized according to the (Dewey) decimal system.*

Desirable: *Which organizational system is used by most of the United States' libraries? (Dewey decimal)*

4. Be sure there is only one correct response.

Undesirable: *Trees which shed their leaves annually are seed-bearing, common.*

Desirable: *Trees which shed their leaves annually are called (deciduous).*

5. Make the blanks of equal length.

Undesirable: *In Greek mythology, Vulcan was the son of (Jupiter) and (Juno).*

Desirable: *In Greek mythology, Vulcan was the son of (Jupiter) and (Juno).*

6. When possible, delete words at the end of the statement after the student has been presented a clearly defined problem.

Undesirable: *(122.5) is the molecular weight of $KClO_3$.*

Desirable: *The molecular weight of $KClO_3$ is (122.5).*

7. Avoid lifting statements directly from the text, lecture or other sources.

8. Limit the required response to a single word or phrase.
-

ESSAY TEST ITEMS

The essay test is probably the most popular of all types of teacher-made tests. In general, a classroom essay test consists of a small number of questions to which the student is expected to demonstrate his/her ability to (a) recall factual knowledge, (b) organize this knowledge and (c) present the knowledge in a logical, integrated answer to the question. An essay test item can be classified as either an extended-response essay item or a short-answer essay item. The latter calls for a more restricted or limited answer in terms of form or scope. An example of each type of essay item follows.

Sample Extended-Response Essay Item

Explain the difference between the S-R (Stimulus-Response) and the S-O-R (Stimulus-Organism-Response) theories of personality. Include in your answer (a) brief descriptions of both theories, (b) supporters of both theories and (c) research methods used to study each of the two theories. (10 pts. 20 minutes)

Sample Short-Answer Essay Item

Identify research methods used to study the S-R (Stimulus-Response) and S-O-R (Stimulus-Organism-Response) theories of personality. (5 pts. 10 minutes)

Advantages in Using Essay Items

Essay items

- are easier and less time consuming to construct than are most other item types.
- provide a means for testing student's ability to compose an answer and present it in a logical manner.
- can efficiently measure higher order cognitive objectives (e.g., analysis, synthesis, evaluation).

Limitations in Using Essay Items

Essay items

- cannot measure a large amount of content or objectives.
- generally provide low test and test scorer reliability.
- require an extensive amount of instructor's time to read and grade.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the grader).

SUGGESTIONS FOR WRITING ESSAY TEST ITEMS

1. Prepare essay items that elicit the type of behavior you want to measure.
Learning Objective: *The student will be able to explain how the normal curve serves as a statistical model.*
Undesirable: *Describe a normal curve in terms of: symmetry, modality, kurtosis and skewness.*
Desirable: *Briefly explain how the normal curve serves as a statistical model for estimation and hypothesis testing.*
2. Phrase each item so that the student's task is clearly indicated.
Undesirable: *Discuss the economic factors which led to the stock market crash of 1929.*
Desirable: *Identify the three major economic conditions which led to the stock market crash of 1929. Discuss briefly each condition in correct chronological sequence and in one paragraph indicate how the three factors were inter-related.*
3. Indicate for each item a point value or weight and an estimated time limit for answering.
Undesirable: *Compare the writings of Bret Harte and Mark Twain in terms of settings, depth of characterization, and dialogue styles of their main characters.*
Desirable: *Compare the writings of Bret Harte and Mark Twain in terms of settings, depth of characterization, and dialogue styles of their main characters. (10 points 20 minutes)*
4. Ask questions that will elicit responses on which experts could agree that one answer is better than another.
5. Avoid giving the student a choice among optional items as this greatly reduces the reliability of the test.
6. It is generally recommended for classroom examinations to administer several short-answer items rather than only one or two extended-response items.

SUGGESTIONS FOR SCORING ESSAY ITEMS

1. Choose a scoring model. Two of the more common scoring models are ANALYTICAL SCORING and GLOBAL QUALITY.
ANALYTICAL SCORING: Each answer is compared to an ideal answer and points are assigned for the inclusion of necessary elements. Grades are based on the number of accumulated points either absolutely (i.e., A=10 or more points, B=6-9 pts., etc.) or relatively (A=top 15% scores, B=next 30% of scores, etc.)
GLOBAL QUALITY: Each answer is read and assigned a score (e.g., grade, total points) based either on the total quality of the response or on the total quality of the response relative to other student answers.

Examples Essay Item and Grading Models

"Americans are a mixed-up people with no sense of ethical values. Everyone knows that baseball is far less necessary than food and steel, yet they pay ball players a lot more than farmers and steelworkers."

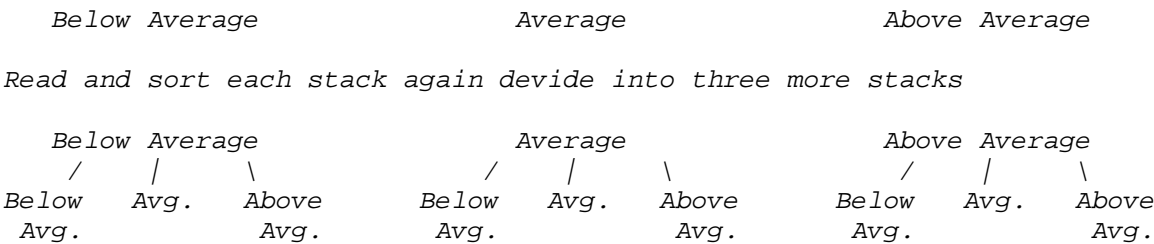
WHY? Use 3-4 sentences to indicate how an economist would explain the above situation.

Analytical Scoring

<u>Necessary Elements to be Included in Response</u>	<u>Points</u>
<i>Salaries are based on demand relative to supply of such services.</i>	3
<i>Excellent ball players are rare.</i>	2
<i>Ball clubs have a high demand for excellent players.</i>	2
<i>Clarity of Response</i>	2
	<hr/>
	9 pts.

Global Quality

Assign scores or grades on the overall quality of the written response as compared to an ideal answer. Or, compare the overall quality of a response to other student responses by sorting the papers into three stacks:



In total, nine discriminations can be used to assign test grades in this manner. The number of stacks or discriminations can vary to meet your needs.

2. Try not to allow factors which are irrelevant to the learning outcomes being measured affect your grading (i.e., handwriting, spelling, neatness).
3. Read and grade all class answers to one item before going on to the next item.
4. Read and grade the answers without looking at the students' names to avoid possible preferential treatment.
5. Occasionally shuffle papers during the reading of answers to help avoid any systematic order

effects (i.e., Sally's "B" work always followed Jim's "A: work thus it looked more like "C" work).

6. When possible, ask another instructor to read and grade your students' responses.

PROBLEM SOLVING TEST ITEMS

Another form of a subjective test item is the problem solving or computational exam question. Such items present the student with a problem situation or task and require a demonstration of work procedures and a correct solution, or just a correct solution. This kind of test item is classified as a subjective type of item due to the procedures used to score item responses. Instructors can assign full or partial credit to either correct or incorrect solutions depending on the quality and kind of work procedures presented. An example of a problem solving test item follows.

Example Problem Solving Test Item

It was calculated that 75 men could complete a strip on a new highway in 70 days. When work was scheduled to commence, it was found necessary to send 25 men on another road project. How many days longer will it take to complete the strip? Show your work for full or partial credit.

Advantages in Using Problem Solving Items

Problem solving items

- minimize guessing by requiring the students to provide an original response rather than to select from several alternatives.
- are easier to construct than are multiple-choice or matching items.
- can most appropriately measure learning objectives which focus on the ability to apply skills or knowledge in the solution of problems.
- can measure an extensive amount of content or objectives.

Limitations in Using Problem Solving Items

Problem solving items

- generally provide low test and test scorer reliability.
- require an extensive amount of instructor time to read and grade.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the grader when partial credit is given).

SUGGESTIONS FOR WRITING PROBLEM SOLVING TEST ITEMS

1. Clearly identify and explain the problem.

Undesirable: *During a car crash, the car slows down at the rate of 490 m/sec². What is the magnitude and direction of the force acting on a 100-kg driver?*

Desirable: *During a car crash, the car slows down at the rate of 490 m/sec². Using the car as a frame of reference, what is the magnitude and direction of the gram*

car as a frame of reference, what is the magnitude and direction of the gram force acting on a 100-kg driver?

2. Provide directions which clearly inform the student of the type of response called for.

Undesirable: *An American tourist in Paris finds that he weighs 70 kilograms. When he left the United States he weighed 144 pounds. What was his net change in weight?*

Desirable: *An American tourist in Paris finds that he weighs 70 kilograms. When he left the United States he weighed 144 pounds. What was his net weight change in pounds?*

3. State in the directions whether or not the student must show his/her work procedures for full or partial credit.

Undesirable: *A double concave lens is made of glass with $n = 1.50$. If the radii of curvature of the two lens surfaces are both 30.0 cm, what is the focal length of the lens?*

Desirable: *A double concave lens is made of glass with $n = 1.50$. If the radii of curvature of the two lens surfaces are both 30.0 cm, what is the focal length of the lens? Show your work to receive full or partial credit.*

4. Clearly separate item parts and indicate their point values.

A man leaves his home and drives to a convention at an average rate of 50 miles per hour. Upon arrival, he finds a telegram advising him to return at once. He catches a plane that takes him back at an average rate of 300 miles per hour.

Undesirable: *If the total traveling time was $1 \frac{3}{4}$ hours, how long did it take him to fly back? How far from his home was the convention?*

Desirable: *If the total traveling time was $1 \frac{3}{4}$ hours:*
(1) How long did it take him to fly back? (1 pt.)
(2) How far from his home was the convention? (1 pt.)
Show your work for full or partial credit.

5. Use figures, conditions and situations which create a realistic problem.

Undesirable: *An automobile weighing 2,840 N (about 640 pounds) is traveling at a speed of 300 miles per hour. What is the car's kinetic energy? Show your work. (2 pts.)*

Desirable: *An automobile weighing 14,200 N (about 3200 pounds) is traveling at a speed of 12m/sec. What is the car's kinetic energy? Show your work. (2 pts.)*

6. Ask questions that elicit responses on which experts could agree that one solution and one or more work procedures are better than others.
7. Work through each problem before classroom administration to double-check accuracy.

PERFORMANCE TEST ITEMS

A performance test item is designed to assess the ability of a student to perform correctly in a simulated situation (i.e., a situation in which the student will be ultimately expected to apply his/her learning). The concept of simulation is central in performance testing; a performance test will simulate to some degree a real life situation to accomplish the assessment. In theory, a performance test could be constructed for any skill and real life situation. In practice, most performance tests have been developed for the assessment of vocational, managerial, administrative, leadership, communication, interpersonal and physical education skills in various simulated situations. An illustrative example of a performance test item is provided below.

Sample Performance Test Item

Assume that some of the instructional objectives of an urban planning course include the development of the student's ability to effectively use the principles covered in the course in various "real life" situations common for an urban planning professional. A performance test item could measure this development by presenting the student with a specific situation which represents a "real life" situation. For example,

An urban planning board makes a last minute request for the professional to act as consultant and critique a written proposal which is to be considered in a board meeting that very evening. The professional arrives before the meeting and has one hour to analyze the written proposal and prepare his critique. The critique presentation is then made verbally during the board meeting; reactions of members of the board or the audience include requests for explanation of specific points or informed attacks on the positions taken by the professional. The performance test designed to simulate this situation would require that the student to be tested role play the professional's part, while students or faculty act the other roles in the situation. Various aspects of the "professional's" performance would than be observed and rated by several judges with the necessary background. The ratings could then be used both to provide the student with a diagnosis of his/her strengths and weaknesses and to contribute to an overall summary evaluation of the student's abilities.

Advantages in Using Performance Test Items

Performance test items

- can most appropriately measure learning objectives which focus on the ability of the students to apply skills or knowledge in real life situations.
- usually provide a degree of test validity not possible with standard paper and pencil test items.
- are useful for measuring learning objectives in the psychomotor domain.

Limitations in Using Performance Test Items

Performance test items

- are difficulty and time consuming to construct.

- are primarily used for testing students individually and not for testing groups. Consequently, they are relatively costly, time consuming, and inconvenient forms of testing.
- generally provide low test and test scorer reliability.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the observer/grader).

SUGGESTIONS FOR WRITING PERFORMANCE TEST ITEMS

1. Prepare items that elicit the type of behavior you want to measure.
 2. Clearly identify and explain the simulated situation to the student.
 3. Make the simulated situation as "life-like" as possible.
 4. Provide directions which clearly inform the students of the type of response called for.
 5. When appropriate, clearly state time and activity limitations in the directions.
 6. Adequately train the observer(s)/scorer(s) to ensure that they are fair in scoring the appropriate behaviors.
-

III. TWO METHODS FOR ASSESSING TEST ITEM QUALITY

This section of the booklet presents two methods for collecting feedback on the quality of your test items. The two methods include using self-review checklists and student evaluation of test item quality. You can use the information gathered from either method to identify strengths and weaknesses in your item writing.

CHECKLIST FOR EVALUATING TEST ITEMS

EVALUATE YOUR TEST ITEMS BY CHECKING THE SUGGESTIONS WHICH YOU FEEL YOU HAVE FOLLOWED.

Multiple-Choice Test Items

- _____ When possible, stated the stem as a direct question rather than as an incomplete statement.
- _____ Presented a definite, explicit and singular question or problem in the stem.
- _____ Eliminated excessive verbiage or irrelevant information from the stem.
- _____ Included in the stem any word(s) that might have otherwise been repeated in each alternative.
- _____ Used negatively stated stems sparingly. When used, underlined and/or capitalized the negative word(s).
- _____ Made all alternatives plausible and attractive to the less knowledgeable or skillful student.
- _____ Made the alternatives grammatically parallel with each other, and consistent with the stem.
- _____ Made the alternatives mutually exclusive.
- _____ When possible, presented alternatives in some logical order (e.g., chronologically, most to least).

- _____ Made sure there was only one correct or best response per item.
- _____ Made alternatives approximately equal in length.
- _____ Avoided irrelevant clues such as grammatical structure, well known verbal associations or connections between stem and answer.
- _____ Used at least four alternatives for each item.
- _____ Randomly distributed the correct response among the alternative positions throughout the test having approximately the same proportion of alternatives a, b, c, d, and e as the correct response.
- _____ Used the alternatives "none of the above" and "all of the above" sparingly. When used, such alternatives were occasionally the correct response.

True-False Test Items

- _____ Based true-false items upon statements that are absolutely true or false, without qualifications or exceptions.
- _____ Expressed the item statement as simply and as clearly as possible.
- _____ Expressed a single idea in each test item.
- _____ Included enough background information and qualifications so that the ability to respond correctly did not depend on some special, uncommon knowledge.
- _____ Avoided lifting statements from the text, lecture or other materials.
- _____ Avoided using negatively stated item statements.
- _____ Avoided the use of unfamiliar language.
- _____ Avoided the use of specific determiners such as "all," "always," "none," "never," etc., and qualifying determiners such as "usually," "sometimes," "often," etc.
- _____ Used more false items than true items (but not more than 15% additional false items).

Matching Test Items

- _____ Included directions which clearly stated the basis for matching the stimuli with the response.
- _____ Explained whether or not a response could be used more than once and indicated where to write the answer.
- _____ Used only homogeneous material.
- _____ When possible, arranged the list of responses in some systematic order (e.g., chronologically, alphabetically).
- _____ Avoided grammatical or other clues to the correct response.
- _____ Kept items brief (limited the list of stimuli to under 10).
- _____ Included more responses than stimuli.
- _____ When possible, reduced the amount of reading time by including only short phrases or single words in the response list.

Completion Test Items

- _____ Omitted only significant words from the statement.
- _____ Did not omit so many words from the statement that the intended meaning was lost.
- _____ Avoided grammatical or other clues to the correct response.
- _____ Included only one correct response per item.
- _____ Made the blanks of equal length.
- _____ When possible, deleted the words at the end of the statement after the student was presented with a clearly defined problem.
- _____ Avoided lifting statements directly from the text, lecture or other sources.
- _____ Limited the required response to a single word or phrase.

Essay Test Items

- _____ Prepared items that elicited the type of behavior you wanted to measure.
- _____ Phrased each item so that the student's task was clearly indicated.
- _____ Indicated for each item a point value or weight and an estimated time limit for answering.
- _____ Asked questions that elicited responses on which experts could agree that one answer is better than others.
- _____ Avoided giving the student a choice among optional items.
- _____ Administered several short-answer items rather than 1 or 2 extended-response items.

Grading Essay Test Items

- _____ Selected an appropriate grading model.
- _____ Tried not to allow factors which were irrelevant to the learning outcomes being measured to affect your grading (e.g., handwriting, spelling, neatness).
- _____ Read and graded all class answers to one item before going on to the next item.
- _____ Read and graded the answers without looking at the student's name to avoid possible preferential treatment.
- _____ Occasionally shuffled papers during the reading of answers.
- _____ When possible, asked another instructor to read and grade your students' responses.

Problem Solving Test Items

- _____ Clearly identified and explained the problem to the student.
- _____ Provided directions which clearly informed the student of the type of response called for.
- _____ Stated in the directions whether or not the student must show work procedures for full or partial credit.
- _____ Clearly separated item parts and indicated their point values.
- _____ Used figures, conditions and situations which created a realistic problem.
- _____ Asked questions that elicited responses on which experts could agree that one solution and one or more work procedures are better than others.

_____ Worked through each problem before classroom administration.

Performance Test Items

_____ Prepared items that elicit the type of behavior you wanted to measure.

_____ Clearly identified and explained the simulated situation to the student.

_____ Made the simulated situation as "life-like" as possible.

_____ Provided directions which clearly inform the students of the type of response called for.

_____ When appropriate, clearly stated time and activity limitations in the directions.

_____ Adequately trained the observer(s)/scorer(s) to ensure that they were fair in scoring the appropriate behaviors.

STUDENT EVALUATION OF TEST ITEM QUALITY**USING ICES QUESTIONNAIRE ITEMS****TO ASSESS YOUR TEST ITEM QUALITY**

The following set of ICES (Instructor and Course Evaluation System) questionnaire items can be used to assess the quality of your test items. The items are presented with their original ICES catalogue number. You are encouraged to include one or more of the items on the ICES evaluation form in order to collect student opinion of your item writing quality.

102:	How would you rate the instructor's examination questions?	Excellent	Poor	116:	Did the exams challenge you to do original thinking?	Yes, very challenging	No, not challenging
103:	How well did examination questions reflect content and emphasis of the course?	Well related	Poorly related	118:	Were there "trick" or trite questions on tests?	Lots of them	Few if any
114:	The exams reflected important points in the reading assignments.	Strongly agree	Strongly disagree	122:	How difficult were the examinations?	Too difficult	Too easy
117:	Examinations mainly tested trivia.	Strongly agree	Strongly disagree	123:	I found I could score reasonably well on exams by just cramming.	Strongly agree	Strongly disagree
119:	Were exam questions worded clearly?	Yes, very clear	No, very unclear	121:	How was the length of exams for the time allotted?	Too long	Too short
115:	Were the instructor's test questions thought provoking?	Definitely yes	Definitely no	109:	Were exams, papers, reports returned with errors explained or personal comments?	Almost always	Almost never
				125:	Were exams adequately discussed upon return?	Yes, adequately	No, not enough

IV. ASSISTANCE OFFERED BY THE OFFICE OF INSTRUCTIONAL RESOURCES (OIR)

The information in the booklet is intended for self-instruction. However, OIR staff members will consult with faculty who wish to analyze and improve their test item writing. The staff can also consult with faculty about other instructional problems. The Measurement and Evaluation Division of OIR also publishes a semi-annual newsletter called Measurement and Evaluation Q & A which discusses various classroom testing and measurement issues. Instructors wishing to receive the newsletter or to acquire OIR assistance can call the Measurement and Evaluation Division at 333-3490.

V. REFERENCES FOR FURTHER READING

- Ebel, Robert L. Measuring educational achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965, Chapters 4-6.
- Ebel, Robert L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, 1972, Chapters 5-8.
- Gronlund, N. E. Measurement and evaluation in teaching. New York: Macmillan Publishing Co., 1976, Chapters 6-9.
- Mehrens, W. A. & Lehmann, I. J. Measurement and evaluation in education and psychology. New York: Holt, Rinehart & Winston, Inc., 1973, Chapters 7-10.
- Nelson, C. H. Measurement and evaluation in the classroom. New York: Macmillan Publishing Co., 1970, Chapters 5-8. Measurement and Evaluation Division, 247 Armory Building. Especially useful for science instruction.
- Payne, David A. The assessment of learning. Lexington, Mass.: D.C. Heath and Co., 1974, Chapters 4-7.
- Scannell, D. P. & Tracy, D. B. Testing and measurement in the classroom. New York: Houghton-Mifflin Co., 1975, Chapters 4-6.
- Thorndike, R. L. (Ed.). Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971, Chapter 9 (Performance testing) and Chapter 10 (Essay exams).
-

<http://www.ucs.umn.edu/oms/multchoice.htmlx>

Last Update: Thu Apr 1 09:46:42 1999

Writing Multiple-Choice Items

[\[Advantages\]](#) [\[Disadvantages\]](#) [\[General Guidelines\]](#)
[\[Item Stem Guidelines\]](#) [\[Item Options Guidelines\]](#) [\[References\]](#)

Multiple-choice (MC) items have many advantages that make them widely used and highly regarded. They also have disadvantages, some of which can be reduced by careful attention to good item-writing and item analysis practice.

Advantages of Multiple-Choice Items

1. Versatility--MC items are adaptable to the measurement of a wide variety of learning outcomes, from knowledge of facts through analysis and interpretation of information to reasoning, making inferences, solving problems, and exercising judgment.
2. Efficiency--Because of the large number of items that can be posed in a given length of time MC items permit wide sampling and broad coverage of the content domain.
3. Scoring accuracy and economy--Expert agreement on the correct answer to MC items is easy to obtain, and scoring keys can be economically applied by machine or clerical assistants.
4. Reliability--Consistency in scoring and wide sampling of content provide test results that are generalizable to the domain of interest.
5. Diagnosis--Patterns of incorrect responses can provide diagnostic information about the learning of individual students or groups.
6. Control of difficulty--The level of difficulty of a test can be increased or decreased by adjusting the degree of similarity among the options to the items.
7. Reduction of guessing--In comparison with two-choice (e.g., true-false) tests, guessing is reduced by MC items.
8. Freedom from response sets--MC items are relatively uninfluenced by response sets, such as a tendency to answer "true."
9. Amenable to item analysis--MC items are amenable to item analysis, by means of which they can be improved.

Disadvantages of Multiple-Choice Items

1. Multiple-choice tests are difficult and time consuming to write. The construction of plausible distractors is especially difficult. The quality of the test is therefore dependent on the item-writing skill of the instructor.

2. There is a tendency to write items requiring only factual knowledge rather than higher level skills and understandings.
3. Performance on MC items can be influenced by student characteristics unrelated to the subject of measurement, such as reading ability, test-wiseness, and risk-taking.
4. MC items are subject to clueing.
5. MC items do not measure ability to organize and express ideas.

General Guidelines

1. Write items that deal with significant facts or concepts, not trivial questions or overly specific details.
2. Write items that have a definite answer. Students may be asked to select either the **correct** answer or the **best** answer. The former instruction is usually more suitable for items dealing with factual knowledge, where the correct answer is a matter of record. For items dealing with interpretation, understanding, or inference, instruction to select the best answer is usually preferred.
3. Communicate clearly. The wording and presentation of the items should not present obstacles to the students' ability to demonstrate what they know. The item should be written in clear language with vocabulary, other than that being tested, as simple and precise as possible.
4. Don't give away the answer by including irrelevant cues in the item.
5. Don't write items that require skills or knowledge irrelevant to what you are trying to measure.
6. Don't use language that may be offensive to some groups.
7. Have items reviewed by knowledgeable persons other than the writer if possible.

Multiple-choice items consist of two parts: the stem and the response options. The stem presents the question or problem. The options include the correct or best answer--the keyed response--and the distractors or foils.

Item Stem Guidelines

1. Write an item as either a direct question (Who was the first President of the United States?) or an incomplete statement (The first President of the United States was _____). Often one form or the other will produce simpler and clearer wording. If not, the question form may be easier for the writer and more straightforward for the student.
2. Present a single, complete problem or question in the stem. Most of the reading should be in the stem.
3. Eliminate excess wording; include only what is necessary to

- present the problem or question.
4. Include in the stem all the information needed to arrive at an unambiguous answer to the item.
 5. Include in the stem any words that would be repeated in each option.
 6. Use an introductory sentence for the item if it seems useful. Two sentences may express the problem more clearly than one.
 7. Write completion items with the blank at the end rather than the beginning or middle.
 8. Avoid the use of negative wording in items. If negatives are necessary, emphasize them with bolding, underlining, or upper case. Do not use negatives in both the stem and the responses, as double-negatives are confusing.
 9. Do not write items that require a series of true-false answers, i.e., questions of the form: "Which of the following is true?"
 10. Make sure that items are independent. The information in one item should not supply the answer to another.
 11. To test understanding and interpretation rather than factual knowledge, ask the questions "How?" and "Why?" rather than "Who?" and "When".
 12. Consider variations on the simple MC format:
 - a. Present material to be interpreted--such as a reading passage, a table, a graph, a map--and base several items on it (being careful about guideline 10, above).
 - b. Use the same set of response options for several items, presenting the options first, in effect creating a small matching task.
 - c. Present items in analogy form: a is to b as 1 is to _____.

Item Options Guidelines

1. Be sure there is one best response to the item. Options must be mutually exclusive and not overlap.
2. Make the length of the options comparable. Avoid overqualifying the keyed response.
3. Make the options parallel in form.
4. Make all options grammatically consistent with the stem.
5. Don't use absolute language, such as "never" and "always" as a means of making options incorrect.
6. Don't repeat key words from the stem in the keyed option.
7. Don't use stereotyped language that may cue the keyed option.
8. Make the distracters plausible and equally attractive to students who do not know the correct response.
9. Use 3-5 options. Four or five options are desirable to reduce guessing, but a good item with three options can be useful. Do not discard an item with only three good options or add implausible options just make the number of options consistent.

10. List the options in a logical order if there is one.
11. Present the options in a list format rather than in a paragraph with the stem.
12. Distribute the correct option randomly among the option positions.
13. Don't make the options overly wordy and confusing.
14. Don't use "All of the above" as an option. "None of the above" should not be used as an option with "best answer" items but can be used effectively with computational items.
15. Sometimes it may be easier to write correct than incorrect options for an item. It is legitimate to ask students to choose the option that is not correct, but heed the caveats regarding negatives in item stem guideline 8, above.
16. It can be helpful to define the class of things to which the correct answer belongs, then write distracters based on members of that class.
17. Consider as distracters responses that are correct but do not answer the question posed by the stem.
18. Obtain distracters from responses of students to items administered in completion or short answer form.

--	--

The University of Minnesota is an equal opportunity educator and employer.

Copyright: © 2003 by the Regents of the University of Minnesota

Page URI: <http://www.ucs.umn.edu/oms/multichoice.htmlx>

Page Coordinator: UCCS [Web Team](#)

[Privacy statement](#)



--

--

<http://ericae.net/pare/getvn.asp?v=4&n=9>

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 1995, EdResearch.org.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Kehoe, Jerard (1995). Writing multiple-choice test items. *Practical Assessment, Research & Evaluation*, 4(9). Available online: <http://edresearch.org/pare/getvn.asp?v=4&n=9>. This paper has been viewed 14,187 times since 11/13/99.

Writing Multiple-Choice Test Items

Jerard Kehoe

Virginia Polytechnic Institute and State University

A notable concern of many teachers is that they frequently have the task of constructing tests but have relatively little training or information to rely on in this task. The objective of this article is to set out some conventional wisdom for the construction of multiple-choice tests, which are one of the most common forms of teacher-constructed tests. The comments which follow are applicable mainly to multiple-choice tests covering fairly broad topic areas.

Before proceeding, it will be useful to establish our terms for discussing multiple-choice items. The *stem* is the introductory question or incomplete statement at the beginning of each item and this is followed by the options. The *options* consist of the answer -- the correct option -- and *distractors*--the incorrect but (we hope) tempting options.

General Objectives

As a rule, one is concerned with writing stems that are clear and parsimonious, answers that are unequivocal and chosen by the students who do best on the test, and distractors that are plausible competitors of the answer as evidenced by the frequency with which they are chosen. Lastly, and probably most important, we should adopt the attitude that items need to be developed over time in the light of evidence that can be obtained from the statistical output typically provided by a measurement services office (where tests are machine-scored) and from "expert" editorial review.

Planning

▶ Find similar papers in

[ERICA Full Text Library](#)
[Pract Assess, Res & Eval](#)
[ERIC RIE & CIJE 1990-](#)
[ERIC On-Demand Docs](#)

▶ Find articles in ERIC written by

[Kehoe, Jerard](#)

The primary objective in planning a test is to outline the actual course content that the test will cover. A convenient way of accomplishing this is to take 10 minutes following each class to list on an index card the important concepts covered in class and in assigned reading for that day. These cards can then be used later as a source of test items. An even more conscientious approach, of course, would be to construct the test items themselves after each class. The advantage of either of these approaches is that the resulting test is likely to be a better representation of course activity than if the test were constructed before the course or after the course, when we usually have only a fond memory or optimistic syllabus to draw from. When we are satisfied that we have an accurate description of the content areas, then all that remains is to construct items that represent specific content areas. In developing good multiple-choice items, three tasks need to be considered: writing stems, writing options, and ongoing item development. The first two are discussed in this article.

Writing Stems

We will first describe some basic rules for the construction of multiple-choice stems, because they are typically, though not necessarily, written before the options.

1. Before writing the stem, identify the one point to be tested by that item. In general, the stem should not pose more than one problem, although the solution to that problem may require more than one step.
2. Construct the stem to be either an incomplete statement or a direct question, avoiding stereotyped phraseology, as rote responses are usually based on verbal stereotypes. For example, the following stems (with answers in parentheses) illustrate undesirable phraseology:

What is the biological theory of recapitulation? (Ontogeny repeats phylogeny)

Who was the chief spokesman for the "American System?" (Henry Clay)

Correctly answering these questions likely depends less on understanding than on recognizing familiar phraseology.

3. Avoid including nonfunctional words that do not contribute to the basis for choosing among the options. Often an introductory statement is included to enhance the appropriateness or significance of an item but does not affect the meaning of the problem in the item. Generally, such superfluous phrases should be excluded. For example, consider:

The American flag has three colors. One of them is (1) red (2) green (3) black

versus

One of the colors of the American flag is (1) red (2) green (3) black

In particular, irrelevant material should not be used to make the answer less obvious. This tends to place too much importance on reading comprehension as a determiner of the correct option.

4. Include as much information in the stem and as little in the options as possible. For example, if the point of an item is to associate a term with its definition, the preferred format would be to present the definition in the stem and several terms as options rather than to present the term in the stem and several definitions as options.

5. Restrict the use of negatives in the stem. Negatives in the stem usually require that the answer be a false statement. Because students are likely in the habit of searching for true statements, this may introduce an unwanted bias.

6. Avoid irrelevant clues to the correct option. Grammatical construction, for example, may lead students to reject options which are grammatically incorrect as the stem is stated. Perhaps more common and subtle, though, is the problem of common elements in the stem and in the answer. Consider the following item:

What led to the formation of the States' Rights Party?

- a. The level of federal taxation*
- b. The demand of states for the right to make their own laws*
- c. The industrialization of the South*
- d. The corruption of federal legislators on the issue of state taxation*

One does not need to know U.S. history in order to be attracted to the answer, b. Other rules that we might list are generally commonsensical, including recommendations for independent and important items and prohibitions against complex, imprecise wording.

Writing Options

Following the construction of the item stem, the likely more difficult task of generating options presents itself. The rules we list below are not likely to simplify this task as much as they are intended to guide our creative efforts.

1. Be satisfied with three or four well constructed options. Generally, the minimal improvement to the item due to that hard-to-come-by fifth option is not worth the effort to construct it. Indeed, all else the same, a test of 10 items each with four

options is likely a better test than a test with nine items of five options each.

2. Construct distractors that are comparable in length, complexity and grammatical form to the answer, avoiding the use of such words as "always," "never," and "all." Adherence to this rule avoids some of the more common sources of biased cueing. For example, we sometimes find ourselves increasing the length and specificity of the answer (relative to distractors) in order to insure its truthfulness. This, however, becomes an easy-to-spot clue for the testwise student. Related to this issue is the question of whether or not test writers should take advantage of these types of cues to construct more tempting distractors. Surely not! The number of students choosing a distractor should depend only on deficits in the content area which the item targets and should not depend on cue biases or reading comprehension differences in "favor" of the distractor.

3. Options which read "none of the above," "both a. and e. above," "all of the above," _etc_, should be avoided when the students have been instructed to choose "the best answer," which implies that the options vary in degree of correctness. On the other hand, "none of the above" is acceptable if the question is factual and is probably desirable if computation yields the answer. "All of the above" is never desirable, as one recognized distractor eliminates it and two recognized answers identify it.

4. After the options are written, vary the location of the answer on as random a basis as possible. A convenient method is to flip two (or three) coins at a time where each possible Head-Tail combination is associated with a particular location for the answer. Furthermore, if the test writer is conscientious enough to randomize the answer locations, students should be informed that the locations are randomized. (Testwise students know that for some instructors the first option is rarely the answer.)

5. If possible, have a colleague with expertise in the content area of the exam review the items for possible ambiguities, redundancies or other structural difficulties. Having completed the items we are typically so relieved that we may be tempted to regard the task as completed and each item in its final and permanent form. Yet, another source of item and test improvement is available to us, namely, statistical analyses of student responses.

This article was adapted with from *Testing Memo 4: Constructing Multiple-Choice Tests -- Part I*, Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

Further Reading

Airasian, P. (1994) *Classroom Assessment*, Second Edition, NY: McGraw-Hill.
Cangelosi, J. (1990) *Designing Tests for Evaluating Student Achievement*. NY: Addison Wellesley.

Grunlund, N (1993) *How to make achievement tests and assessments*, 5th edition, NY: Allen and Bacon.

Haladyna, T.M. & Downing, S.M. (1989) Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2 (1), 51-78.

Descriptors: *Culture Fair Tests; *Distractors (Tests); Educational Assessment; Item Bias; Measurement Techniques; *Multiple Choice Tests; Scoring; *Statistical Analysis; Stereotypes; *Test Construction; Test Items; Test Theory

<http://ericae.net/pare/getvn.asp?v=4&n=11>

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Search:

Copyright 1995, EdResearch.org.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Frary, Robert B. (1995). More multiple-choice item writing do's and don'ts. *Practical Assessment, Research & Evaluation*, 4(11). Available online: <http://edresearch.org/pare/getvn.asp?v=4&n=11>. This paper has been viewed 12,588 times since 11/13/99.

More Multiple-choice Item Writing Do's And Don'ts

Robert B. Frary
Virginia Polytechnic Institute and State University

▶ Find similar papers in

[ERICA Full Text Library](#)
[Pract Assess, Res & Eval](#)
[ERIC RIE & CIJE 1990-](#)
[ERIC On-Demand Docs](#)

▶ Find articles in ERIC written by

[Frary, Robert B.](#)

Kehoe(1995) gave a few suggestions for item-writing, but only to a limited extent, due to its coverage of other aspects of test development. What follows here is a fairly comprehensive list of recommendations for writing multiple choice items. Some of these are backed up by psychometric research; i.e., it has been found that, generally, the resulting scores are more accurate indicators of each student's knowledge when the recommendations are followed than when they are violated. Other recommendations result from logical deduction.

Content

1. Do ask questions that require more than knowledge of facts. For example, a question might require selection of the best answer when all of the options contain elements of correctness. Such questions tend to be more difficult and discriminating than questions that merely ask for a fact. Justifying the "bestness" of the keyed option may be as challenging to the instructor as the item was to the students, but, after all, isn't challenging students and responding to their challenges a big part of what being a teacher is all about?

2. Don't offer superfluous information as an introduction to a question, for example, "*The presence and association of the male seems to have profound effects on female physiology in domestic animals. Research has shown that in cattle presence of a bull has the following effect:*" This approach probably represents an unconscious effort to

continue teaching while testing and is not likely to be appreciated by the students, who would prefer direct questions and less to read. The stem just quoted could be condensed to "Research has shown that the presence of a bull has which of the following effects on cows?" (17 words versus 30).

Structure

3. Don't ask a question that begins, "*Which of the following is true [or false]?*" followed by a collection of unrelated options. Each test question should focus on some specific aspect of the course. Therefore, it's OK to use items that begin, "Which of the following is true [or false] concerning X?" followed by options all pertaining to X. However, this construction should be used sparingly if there is a tendency to resort to trivial reasons for falseness or an opposite tendency to offer options that are too obviously true. A few true-false questions (in among the multiple-choice questions) may forestall these problems. The options would be: *1) True 2) False*.

4. Don't use items like the following:
What is (are) the capital(s) of Bolivia?

A. La Paz B. Sucre C. Santa Cruz

1) A only 4) Both A and B

2) B only 5) All of the above

3) C only

Research on this item type has consistently shown it to be easier and less discriminating than items with distinct options. In the example above, one only needs to remember that Bolivia has two capitals to be assured of answering correctly. This problem can be alleviated by offering all possible combinations of the three basic options, namely:

1) A only, 2) B only, 3) C only, 4) A and B, 5) A and C, 6) B and C, 7) A, B, and C, 8) None of the above.

However, due to its complexity, initial use of this adaptation should be limited.

Options

5. Do ask questions with varying numbers of options. There is no psychometric advantage to having a uniform number, especially if doing so results in options that are so implausible that no one or almost no one marks them. In fact, some valid and important questions demand only two or three options, e.g., "*If drug X is administered, body temperature will probably: 1) increase, 2) stay about the same, 3) decrease.*"

6. Don't put negative options following a negative stem. Empirically (or statistically) such items may appear to perform adequately, but this is probably only because brighter students who naturally tend to get higher scores are also better able to cope with the logical complexity of a double negative.

7. Don't use "*all of the above*." Recognition of one wrong option eliminates "all of the above," and recognition of two right options identifies it as the answer, even if the other options are completely unknown to the student. Probably some instructors use items with "all of the above" as yet another way of extending their teaching into the test (see 2 above). It just seems so good to have the students affirm, say, all of the major causes of some phenomenon. With this approach, "all of the above" is the answer to almost every item containing it, and the students soon figure this out.

8. Do ask questions with "*none of the above*" as the final option, especially if the answer requires computation. Its use makes the question harder and more discriminating, because the uncertain student cannot focus on a set of options that must contain the answer. Of course, "*none of the above*" cannot be used if the question requires selection of the best answer and should not be used following a negative stem. Also, it is important that "*none of the above*" should be the answer to a reasonable proportion of the questions containing it.

9. Don't include superfluous information in the options. The reasons given for 8 above apply. In addition, as another manifestation of the desire to teach while testing, the additional information is likely to appear on the correct answer: 1) *W*, 2) *X*, 3) *Y*, *because*, 4) *Z*. Students are very sensitive to this tendency and take advantage of it.

10. Don't use specific determiners in distractors. Sometimes in a desperate effort to produce another, often unneeded, distractor (see 5 above), a statement is made incorrect by the inclusion of words like all or never, e.g., "*All humans have 46 chromosomes*." Students learn to classify such statements as distractors when otherwise ignorant.

11. Don't repeat wording from the stem in the correct option. Again, an ignorant student will take advantage of this practice.

Errors to avoid

Most violations of the recommendations given thus far should not be classified as outright errors, but, instead, perhaps, as lapses of judgement. And, as almost all rules have exceptions, there are probably circumstances where some of 1-11 above would not hold. However, there are three not-too-common item-writing/test-preparation errors that represent nothing less than negligence. They are now mentioned to encourage careful preparation and proofreading of tests:

Typos. These are more likely to appear in distractors than in the stem and the correct answer, which get more scrutiny from the test preparer. Students easily become aware of this tendency if it is present.

Grammatical inconsistency between stem and options. Almost always, the stem and the correct answer are grammatically consistent, but distractors, often produced as afterthoughts, may not mesh properly with the stem. Again, students quickly learn to take advantage of this foible.

Overlapping distractors. For example: *Due to budget cutbacks, the university library now subscribes to fewer than _?_ periodicals. 1) 25,000 2) 20,000 3) 15,000 4) 10,000*

Perhaps surprisingly, not all students "catch on" to items like this, but many do. Worse yet, the instructor might indicate option 2 as the correct answer. Finally, we consider an item-writing foible reported by Smith (1982). What option would you select among the following (stem omitted)?

1) *Abraham Lincoln* 3) *Stephen A. Douglas*
2) *Robert E. Lee* 4) *Andrew Jackson*

The testwise but ignorant student will select Lincoln because it represents the intersection of two categories of prominent nineteenth century people, namely, presidents and men associated with the Civil War.

Try this one:

1) *before breakfast* 3) *on a full stomach*
2) *with meals* 4) *before going to bed*

Three options have to do with eating, and two with the time of day. Only one relates to both. Unfortunately, some item writers consciously or unconsciously construct items of this type with the intersection invariably the correct answer.

This article was adapted from *Testing Memo 10: Some Multiple-choice Item Writing Do's And Don'ts*, Office of Measurement and Research Services, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060

Further Reading

- Airasian, P. (1994). *Classroom Assessment*, Second Edition, NY: McGraw-Hill.
- Brown, F. (1983). *Principles of Educational and Psychological Testing*, Third edition, NY: Holt Rinehart, Winston. Chapter 11.
- Cangelosi, J. (1990). *Designing Tests for Evaluating Student Achievement*. NY: Longman.
- Grunlund, N (1993). *How to make achievement tests and assessments*, 5th edition, NY: Allen and Bacon.
- Haladyna, T.M. & Downing, S.M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, 2 (1), 51-78.
- Kehoe, J (1995). Writing Multiple-Choice Test Items. *Practical Assessment, Research and Evaluation*, 4(4). [Available online <http://ericae.net/pare/getvn.asp?v4&n4>].
- Roid, G.H. & Haladvna, T.M. (1980). The emergence of an item writing technology.

Review of Educational Research, 49, 252-279.

Smith, J. K. (1982). Converging on correct answers: A peculiarity of multiple-choice items. *Journal of Educational Measurement*, 19, 211-220.

Wesman, A.G. (1971). Writing the test item. In R.L. Thorndike (Ed.) *Educational Measurement* (1st ed, pp 99-111). Washington, DC: American Council on Education

Descriptors: Culture Fair Tests; *Distractors (Tests); Educational Assessment; Item Bias; Measurement Techniques; *Multiple Choice Tests; *Psychometrics; Scoring; *Statistical Analysis; Stereotypes; *Test Construction; Test Items; Test Theory

APPENDIX H:

Articles: ["Scoring Rubrics: What, When and How"](#)
["Designing Schoring Rubrics for Your Classroom"](#)

<http://ericae.net/pare/getvn.asp?v=7&n=3>

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2000, EdResearch.org.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Moskal, Barbara M. (2000). Scoring rubrics: what, when and how?. *Practical Assessment, Research & Evaluation*, 7(3). Available online: <http://edresearch.org/pare/getvn.asp?v=7&n=3>. This paper has been viewed 51,287 times since 3/29/00.

Scoring Rubrics: What, When and How?

Barbara M. Moskal

Associate Director of the Center for Engineering Education
Assistant Professor of Mathematical and Computer Sciences
Colorado School of Mines

Scoring rubrics have become a common method for evaluating student work in both the K-12 and the college classrooms. The purpose of this paper is to describe the different types of scoring rubrics, explain why scoring rubrics are useful and provide a process for developing scoring rubrics. This paper concludes with a description of resources that contain examples of the different types of scoring rubrics and further guidance in the development process.

What is a scoring rubric?

Scoring rubrics are descriptive scoring schemes that are developed by teachers or other evaluators to guide the analysis of the products or processes of students' efforts (Brookhart, 1999). Scoring rubrics are typically employed when a judgement of quality is required and may be used to evaluate a broad range of subjects and activities. One common use of scoring rubrics is to guide the evaluation of writing samples. Judgements concerning the quality of a given writing sample may vary depending upon the criteria established by the individual evaluator. One evaluator may heavily weigh the evaluation process upon the linguistic structure, while another evaluator may be more interested in the persuasiveness of the argument. A high quality essay is likely to have a combination of these and other factors. By developing a pre-defined scheme for the evaluation process, the subjectivity involved in evaluating an essay becomes more objective.

Figure 1 displays a scoring rubric that was developed to guide the evaluation of student writing samples in a college classroom (based loosely on Leydens & Thompson, 1997). This is an example of a holistic scoring rubric with four score levels. Holistic rubrics will be discussed in detail later in this document. As the example illustrates, each score category

▶ Find similar papers in

[ERICA Full Text Library](#)
[Pract Assess, Res & Eval](#)
[ERIC RIE & CIJE 1990-](#)
[ERIC On-Demand Docs](#)

▶ Find articles in ERIC written by

[Moskal, Barbara M.](#)

describes the characteristics of a response that would receive the respective score. By having a description of the characteristics of responses within each score category, the likelihood that two independent evaluators would assign the same score to a given response is increased. This concept of examining the extent to which two independent evaluators assign the same score to a given response is referred to as "rater reliability."

Figure 1.

Example of a scoring rubric designed to evaluate college writing samples.

<p style="text-align: center;">-3-</p> <p style="text-align: center;">Meets Expectations for a first Draft of a Professional Report</p> <ul style="list-style-type: none">• The document can be easily followed. A combination of the following are apparent in the document:<ol style="list-style-type: none">1. Effective transitions are used throughout,2. A professional format is used,3. The graphics are descriptive and clearly support the document's purpose.• The document is clear and concise and appropriate grammar is used throughout.
<p style="text-align: center;">-2-</p> <p style="text-align: center;">Adequate</p> <ul style="list-style-type: none">• The document can be easily followed. A combination of the following are apparent in the document:<ol style="list-style-type: none">1. Basic transitions are used,2. A structured format is used,3. Some supporting graphics are provided, but are not clearly explained.• The document contains minimal distractions that appear in a combination of the following forms:<ol style="list-style-type: none">1. Flow in thought2. Graphical presentations3. Grammar/mechanics
<p style="text-align: center;">-1-</p> <p style="text-align: center;">Needs Improvement</p> <ul style="list-style-type: none">• Organization of document is difficult to follow due to a combination of following:

1. Inadequate transitions
 2. Rambling format
 3. Insufficient or irrelevant information
 4. Ambiguous graphics
- The document contains numerous distractions that appear in the a combination of the following forms:
 1. Flow in thought
 2. Graphical presentations
 3. Grammar/mechanics

-0-
Inadequate

- There appears to be no organization of the document's contents.
- Sentences are difficult to read and understand.

When are scoring rubrics an appropriate evaluation technique?

Writing samples are just one example of performances that may be evaluated using scoring rubrics. Scoring rubrics have also been used to evaluate group activities, extended projects and oral presentations (e.g., Chicago Public Schools, 1999; Danielson, 1997a; 1997b; Schrock, 2000; Moskal, 2000). They are equally appropriate to the English, Mathematics and Science classrooms (e.g., Chicago Public Schools, 1999; State of Colorado, 1999; Danielson, 1997a; 1997b; Danielson & Marquez, 1998; Schrock, 2000). Both pre-college and college instructors use scoring rubrics for classroom evaluation purposes (e.g., State of Colorado, 1999; Schrock, 2000; Moskal, 2000; Knecht, Moskal & Pavelich, 2000). Where and when a scoring rubric is used does not depend on the grade level or subject, but rather on the purpose of the assessment.

Scoring rubrics are one of many alternatives available for evaluating student work. For example, checklists may be used rather than scoring rubrics in the evaluation of writing samples. Checklists are an appropriate choice for evaluation when the information that is sought is limited to the determination of whether specific criteria have been met. Scoring rubrics are based on descriptive scales and support the evaluation of the extent to which criteria has been met.

The assignment of numerical weights to sub-skills within a process is another evaluation technique that may be used to determine the extent to which given criteria has been met. Numerical values, however, do not provide students with an indication as to how to improve their performance. A student who receives a "70" out of "100". may not know how to improve

his or her performance on the next assignment. Scoring rubrics respond to this concern by providing descriptions at each level as to what is expected. These descriptions assist the students in understanding why they received the score that they did and what they need to do to improve their future performances.

Whether a scoring rubric is an appropriate evaluation technique is dependent upon the purpose of the assessment. Scoring rubrics provide at least two benefits in the evaluation process. First, they support the examination of the extent to which the specified criteria has been reached. Second, they provide feedback to students concerning how to improve their performances. If these benefits are consistent with the purpose of the assessment, then a scoring rubric is likely to be an appropriate evaluation technique.

What are the different types of scoring rubrics?

Several different types of scoring rubrics are available. Which variation of the scoring rubric should be used in a given evaluation is also dependent upon the purpose of the evaluation. This section describes the differences between analytic and holistic scoring rubrics and between task specific and general scoring rubrics.

Analytic versus Holistic

In the initial phases of developing a scoring rubric, the evaluator needs to determine what will be the evaluation criteria. For example, two factors that may be considered in the evaluation of a writing sample are whether appropriate grammar is used and the extent to which the given argument is persuasive. An analytic scoring rubric, much like the checklist, allows for the separate evaluation of each of these factors. Each criterion is scored on a different descriptive scale (Brookhart, 1999).

The rubric that is displayed in Figure 1 could be extended to include a separate set of criteria for the evaluation of the persuasiveness of the argument. This extension would result in an analytic scoring rubric with two factors, quality of written expression and persuasiveness of the argument. Each factor would receive a separate score. Occasionally, numerical weights are assigned to the evaluation of each criterion. As discussed earlier, the benefit of using a scoring rubric rather than weighted scores is that scoring rubrics provide a description of what is expected at each score level. Students may use this information to improve their future performance.

Occasionally, it is not possible to separate an evaluation into independent factors. When there is an overlap between the criteria set for the evaluation of the different factors, a holistic scoring rubric may be preferable to an analytic scoring rubric. In a holistic scoring rubric, the criteria is considered in combination on a single descriptive scale (Brookhart, 1999). Holistic scoring rubrics support broader judgments concerning the quality of the process or the product.

Selecting to use an analytic scoring rubric does not eliminate the possibility of a holistic

factor. A holistic judgment may be built into an analytic scoring rubric as one of the score categories. One difficulty with this approach is that overlap between the criteria that is set for the holistic judgment and the other evaluated factors cannot be avoided. When one of the purposes of the evaluation is to assign a grade, this overlap should be carefully considered and controlled. The evaluator should determine whether the overlap is resulting in certain criteria are being weighted more than was originally intended. In other words, the evaluator needs to be careful that the student is not unintentionally severely penalized for a given mistake.

General versus Task Specific

Scoring rubrics may be designed for the evaluation of a specific task or the evaluation of a broader category of tasks. If the purpose of a given course is to develop a student's oral communication skills, a general scoring rubric may be developed and used to evaluate each of the oral presentations given by that student. This approach would allow the students to use the feedback that they acquired from the last presentation to improve their performance on the next presentation.

If each oral presentation focuses upon a different historical event and the purpose of the assessment is to evaluate the students' knowledge of the given event, a general scoring rubric for evaluating a sequence of presentations may not be adequate. Historical events differ in both influencing factors and outcomes. In order to evaluate the students' factual and conceptual knowledge of these events, it may be necessary to develop separate scoring rubrics for each presentation. A "Task Specific" scoring rubric is designed to evaluate student performances on a single assessment event.

Scoring rubrics may be designed to contain both general and task specific components. If the purpose of a presentation is to evaluate students' oral presentation skills and their knowledge of the historical event that is being discussed, an analytic rubric could be used that contains both a general component and a task specific component. The oral component of the rubric may consist of a general set of criteria developed for the evaluation of oral presentations; the task specific component of the rubric may contain a set of criteria developed with the specific historical event in mind.

How are scoring rubrics developed?

The first step in developing a scoring rubric is to clearly identify the qualities that need to be displayed in a student's work to demonstrate proficient performance (Brookhart, 1999). The identified qualities will form the top level or levels of scoring criteria for the scoring rubric. The decision can then be made as to whether the information that is desired from the evaluation can best be acquired through the use of an analytic or holistic scoring rubric. If an analytic scoring rubric is created, then each criterion is considered separately as the descriptions of the different score levels are developed. This process results in separate descriptive scoring schemes for each evaluation factor. For holistic scoring rubrics, the collection of criteria is considered throughout the construction of each level of the scoring

rubric and the result is a single descriptive scoring scheme.

After defining the criteria for the top level of performance, the evaluator's attention may be turned to defining the criteria for lowest level of performance. What type of performance would suggest a very limited understanding of the concepts that are being assessed? The contrast between the criteria for top level performance and bottom level performance is likely to suggest appropriate criteria for middle level of performance. This approach would result in three score levels.

If greater distinctions are desired, then comparisons can be made between the criteria for each existing score level. The contrast between levels is likely to suggest criteria that may be used to create score levels that fall between the existing score levels. This comparison process can be used until the desired number of score levels is reached or until no further distinctions can be made. If meaningful distinctions between the score categories cannot be made, then additional score categories should not be created (Brookhart, 1999). It is better to have a few meaningful score categories than to have many score categories that are difficult or impossible to distinguish.

Each score category should be defined using descriptions of the work rather than judgements about the work (Brookhart, 1999). For example, "Student's mathematical calculations contain no errors," is preferable over, "Student's calculations are good." The phrase "are good" requires the evaluator to make a judgement whereas the phrase "no errors" is quantifiable. In order to determine whether a rubric provides adequate descriptions, another teacher may be asked to use the scoring rubric to evaluate a sub-set of student responses. Differences between the scores assigned by the original rubric developer and the second scorer will suggest how the rubric may be further clarified.

Resources

Currently, there is a broad range of resources available to teachers who wish to use scoring rubrics in their classrooms. These resources differ both in the subject that they cover and the level that they are designed to assess. The examples provided below are only a small sample of the information that is available.

For K-12 teachers, the State of Colorado (1998) has developed an on-line set of general, holistic scoring rubrics that are designed for the evaluation of various writing assessments. The Chicago Public Schools (1999) maintain an extensive electronic list of analytic and holistic scoring rubrics that span the broad array of subjects represented throughout K-12 education. For mathematics teachers, Danielson has developed a collection of reference books that contain scoring rubrics that are appropriate to the elementary, middle school and high school mathematics classrooms (1997a, 1997b; Danielson & Marquez, 1998).

Resources are also available to assist college instructors who are interested in developing and using scoring rubrics in their classrooms. *Kathy Schrock's Guide for Educators* (2000) contains electronic materials for both the pre-college and the college classroom. In *The Art*

and Science of Classroom Assessment: The Missing Part of Pedagogy, Brookhart (1999) provides a brief, but comprehensive review of the literature on assessment in the college classroom. This includes a description of scoring rubrics and why their use is increasing in the college classroom. Moskal (1999) has developed a web site that contains links to a variety of college assessment resources, including scoring rubrics.

The resources described above represent only a fraction of those that are available. The ERIC Clearinghouse on Assessment and Evaluation [ERIC/AE] provides several additional useful web sites. One of these, *Scoring Rubrics - Definitions & Constructions* (2000b), specifically addresses questions that are frequently asked with regard to scoring rubrics. This site also provides electronic links to web resources and bibliographic references to books and articles that discuss scoring rubrics. For more recent developments within assessment and evaluation, a search can be completed on the abstracts of papers that will soon be available through ERIC/AE (2000a). This site also contains a direct link to ERIC/AE abstracts that are specific to scoring rubrics.

Search engines that are available on the web may be used to locate additional electronic resources. When using this approach, the search criteria should be as specific as possible. Generic searches that use the terms "rubrics" or "scoring rubrics" will yield a large volume of references. When seeking information on scoring rubrics from the web, it is advisable to use an advanced search and specify the grade level, subject area and topic of interest. If more resources are desired than result from this conservative approach, the search criteria can be expanded.

References

- Brookhart, S. M. (1999). *The Art and Science of Classroom Assessment: The Missing Part of Pedagogy*. ASHE-ERIC Higher Education Report (Vol. 27, No.1). Washington, DC: The George Washington University, Graduate School of Education and Human Development.
- Chicago Public Schools (1999). *Rubric Bank*. [Available online at: http://intranet.cps.k12.il.us/Assessments/Ideas_and_Rubrics/Rubric_Bank/rubric_bank.html].
- Danielson, C. (1997a). *A Collection of Performance Tasks and Rubrics: Middle School Mathematics*. Larchmont, NY: Eye on Education Inc.
- Danielson, C. (1997b). *A Collection of Performance Tasks and Rubrics: Upper Elementary School Mathematics*. Larchmont, NY: Eye on Education Inc.
- Danielson, C. & Marquez, E. (1998). *A Collection of Performance Tasks and Rubrics: High School Mathematics*. Larchmont, NY: Eye on Education Inc.
- ERIC/AE (2000a). *Search ERIC/AE draft abstracts*. [Available online at: <http://ericae.net/sinprog.htm>].
- ERIC/AE (2000b). *Scoring Rubrics - Definitions & Construction* [Available online at: http://ericae.net/faqs/rubrics/scoring_rubrics.htm].
- Knecht, R., Moskal, B. & Pavelich, M. (2000). *The Design Report Rubric: Measuring and Tracking Growth through Success*. Paper to be presented at the annual meeting of

the American Society for Engineering Education.

Leydens, J. & Thompson, D. (August, 1997), *Writing Rubrics Design (EPICS) I*, Internal Communication, Design (EPICS) Program, Colorado School of Mines.

Moskal, B. (2000). *Assessment Resource Page*. [Available online at:

<http://www.mines.edu/Academic/assess/Resource.htm>].

Schrock, K. (2000). *Kathy Schrock's Guide for Educators*. [Available online at:

<http://school.discovery.com/schrockguide/assess.html>].

State of Colorado (1998). The Rubric. [Available online at:

<http://www.cde.state.co.us/cdedepcom/asrubric.htm#writing>].

Descriptors: *Rubrics; Scoring; *Student Evaluation; *Test Construction; *Evaluation Methods; Grades; Grading; *Scoring

<http://ericae.net/pare/getvn.asp?v=7&n=25>

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal. ISSN 1531-7714

Copyright 2001, EdResearch.org.

Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. Please notify the editor if an article is to be used in a newsletter.

Mertler, Craig A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). Available online: <http://edresearch.org/pare/getvn.asp?v=7&n=25>. This paper has been viewed 15,462 times since 12/11/01.

Designing Scoring Rubrics for Your Classroom

[Craig A. Mertler](#)

Bowling Green State University

Find similar papers in

[ERICA Full Text Library](#)
[Pract Assess. Res & Eval](#)
[ERIC RIE & CIJE 1990-](#)
[ERIC On-Demand Docs](#)

Find articles in ERIC written by

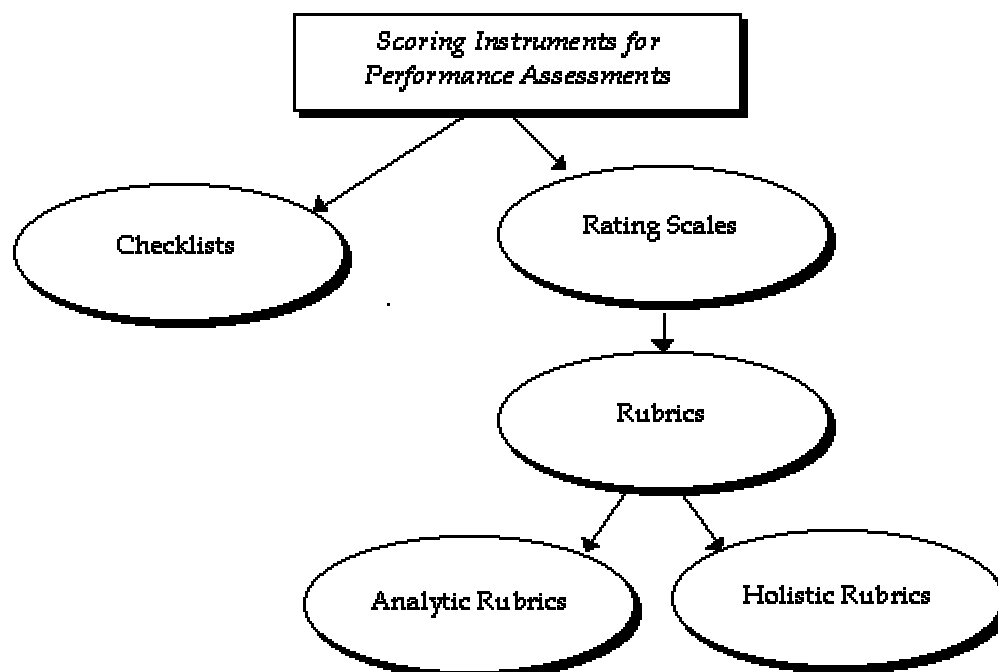
[Mertler, Craig A.](#)

Rubrics are rating scales-as opposed to checklists-that are used with performance assessments. They are formally defined as scoring guides, consisting of specific pre-established performance criteria, used in evaluating student work on performance assessments. Rubrics are typically the specific form of scoring instrument used when evaluating student performances or products resulting from a performance task.

There are two types of rubrics: holistic and analytic (see Figure 1). A **holistic rubric** requires the teacher to score the overall process or product as a whole, without judging the component parts separately (Nitko, 2001). In contrast, with an **analytic rubric**, the teacher scores separate, individual parts of the product or performance first, then sums the individual scores to obtain a total score (Moskal, 2000; Nitko, 2001).

Figure 1:

Types of scoring instruments for performance assessments



Holistic rubrics are customarily utilized when errors in some part of the process can be tolerated provided the overall quality is high (Chase, 1999). Nitko (2001) further states that use of holistic rubrics is probably more appropriate when performance tasks require students to create some sort of response and where there is no definitive correct answer. The focus of a score reported using a holistic rubric is on the overall quality, proficiency, or understanding of the specific content and skills-it involves assessment on a unidimensional level (Mertler, 2001). Use of holistic rubrics can result in a somewhat quicker scoring process than use of analytic rubrics (Nitko, 2001). This is basically due to the fact that the teacher is required to read through or otherwise examine the student product or performance only once, in order to get an "overall" sense of what the student was able to accomplish (Mertler, 2001). Since assessment of the overall performance is the key, holistic rubrics are also typically, though not exclusively, used when the purpose of the performance assessment is summative in nature. At most, only limited feedback is provided to the student as a result of scoring performance tasks in this manner. A template for holistic scoring rubrics is presented in Table 1.

Table 1: <i>Template for Holistic Rubrics</i>	
<u>Score</u>	<u>Description</u>
5	Demonstrates complete understanding of the problem. All requirements of task are included in response.
4	Demonstrates considerable understanding of the problem. All requirements of task are included.
3	Demonstrates partial understanding of the problem. Most requirements of task are included.
2	Demonstrates little understanding of the problem. Many requirements of task are missing.
1	Demonstrates no understanding of the problem.
0	No response/task not attempted.

Analytic rubrics are usually preferred when a fairly focused type of response is required (Nitko, 2001); that is, for performance tasks in which there may be one or two acceptable responses and creativity is not an essential feature of the students' responses. Furthermore, analytic rubrics result initially in several scores, followed by a summed total score-their use represents assessment on a multidimensional level (Mertler, 2001). As previously mentioned, the use of analytic rubrics can cause the scoring process to be substantially slower, mainly because assessing several different skills or characteristics individually requires a teacher to examine the product several times. Both their construction and use can be quite time-consuming. A general rule of thumb is that an individual's work should be examined a separate time for each of the specific performance tasks or scoring criteria (Mertler, 2001). However, the advantage to the use of analytic rubrics is quite substantial. The degree of feedback offered to students-and to teachers-is significant. Students receive specific feedback on their performance with respect to each of the individual scoring criteria-something that does not happen when using holistic rubrics (Nitko, 2001). It is possible to then create a "profile" of specific student strengths and weaknesses (Mertler,

2001). A template for analytic scoring rubrics is presented in Table 2.

Table 2: <i>Template for analytic rubrics</i>					
	Beginning 1	Developing 2	Accomplished 3	Exemplary 4	Score
Criteria #1	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #2	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #3	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	
Criteria #4	Description reflecting beginning level of performance	Description reflecting movement toward mastery level of performance	Description reflecting achievement of mastery level of performance	Description reflecting highest level of performance	

Prior to designing a specific rubric, a teacher must decide whether the performance or product will be scored holistically or analytically (Airasian, 2000 & 2001). Regardless of which type of rubric is selected, specific performance criteria and observable indicators must be identified as an initial step to development. The decision regarding the use of a holistic or analytic approach to scoring has several possible implications. The most important of these is that teachers must consider first how they intend to use the results. If an overall, summative score is desired, a holistic scoring approach would be more desirable. In contrast, if formative feedback is the goal, an analytic scoring rubric should be used. It is important to note that one type of rubric is not inherently better than the other—you must find a format that works best for your purposes (Montgomery, 2001). Other implications include the time requirements, the nature of the task itself, and the specific performance criteria being observed.

As you saw demonstrated in the templates (Tables 1 and 2), the various levels of student performance can be defined using either quantitative (i.e., numerical) or qualitative (i.e., descriptive) labels. In some instances, teachers might want to utilize both quantitative and qualitative labels. If a rubric contains four levels of proficiency or understanding on a continuum, quantitative labels would typically range from "1" to "4." When using qualitative labels, teachers have much more flexibility, and can be more creative. A common type of qualitative scale might include the following labels: master, expert, apprentice, and novice. Nearly any type of qualitative scale will suffice, provided it "fits" with the task. One potentially frustrating aspect of scoring student work with rubrics is the issue of

somehow converting them to "grades." It is not a good idea to think of rubrics in terms of percentages (Trice, 2000). For example, if a rubric has six levels (or "points"), a score of 3 should not be equated to 50% (an "F" in most letter grading systems). The process of converting rubric scores to grades or categories is more a process of logic than it is a mathematical one. Trice (2000) suggests that in a rubric scoring system, there are typically more scores at the average and above average categories (i.e., equating to grades of "C" or better) than there are below average categories. For instance, if a rubric consisted of nine score categories, the equivalent grades and categories might look like this:

Table 3: <i>Sample grades and categories</i>		
<i>Rubric Score</i>	<i>Grade</i>	<i>Category</i>
8	A+	Excellent
7	A	Excellent
6	B+	Good
5	B	Good
4	C+	Fair
3	C	Fair
2	U	Unsatisfactory
1	U	Unsatisfactory
0	U	Unsatisfactory

When converting rubric scores to grades (typical at the secondary level) or descriptive feedback (typical at the elementary level), it is important to remember that there is not necessarily one correct way to accomplish this. The bottom line for classroom teachers is that they must find a system of conversion that works for them and fits comfortably into their individual system of reporting student performance.

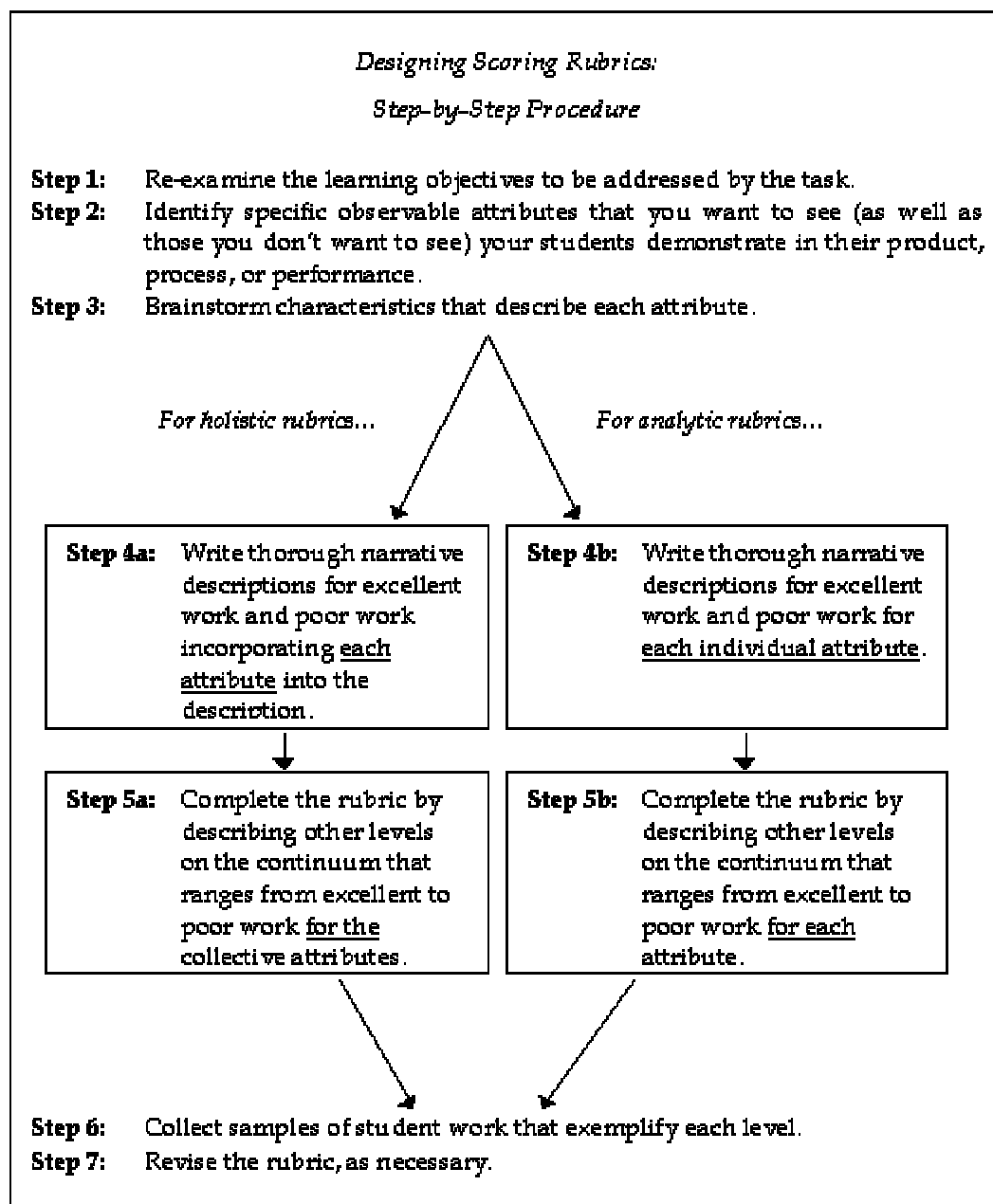
Steps in the Design of Scoring Rubrics

A step-by-step process for designing scoring rubrics for classroom use is presented below. Information for these procedures was compiled from various sources (Airasian, 2000 & 2001; Mertler, 2001; Montgomery, 2001; Nitko, 2001; Tombari & Borich, 1999). The steps will be summarized and discussed, followed by presentations of two sample scoring rubrics.

- Step 1:** *Re-examine the learning objectives to be addressed by the task.* This allows you to match your scoring guide with your objectives and actual instruction.
- Step 2:** *Identify specific **observable** attributes that you want to see (as well as those you don't want to see) your students demonstrate in their product, process, or performance.* Specify the characteristics, skills, or behaviors that you will be looking for, as well as common mistakes you do not want to see.
- Step 3:** *Brainstorm characteristics that describe each attribute.* Identify ways to describe above average, average, and below average performance for each observable attribute identified in Step 2.
- Step 4a:** *For holistic rubrics, write thorough narrative descriptions for excellent work and poor work incorporating each attribute into the description.* Describe the highest and lowest levels of performance combining the descriptors for all attributes.
- Step 4b:** *For analytic rubrics, write thorough narrative descriptions for excellent work and poor work for each individual attribute.* Describe the highest and lowest levels of performance using the descriptors for each attribute separately.
- Step 5a:** *For holistic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for the collective attributes.* Write descriptions for all intermediate levels of performance.
- Step 5b:** *For analytic rubrics, complete the rubric by describing other levels on the continuum that ranges from excellent to poor work for each attribute.* Write descriptions for all intermediate levels of performance for each attribute separately.
- Step 6:** *Collect samples of student work that exemplify each level.* These will help you score in the future by serving as benchmarks.
- Step 7:** *Revise the rubric, as necessary.* Be prepared to reflect on the effectiveness of the rubric and revise it prior to its next implementation.

These steps involved in the design of rubrics have been summarized in Figure 2.

Figure 2:
Designing Scoring Rubrics: Step-by-step procedures



Two Examples

Two sample scoring rubrics corresponding to specific performance assessment tasks are presented next. Brief discussions precede the actual rubrics. For illustrative purposes, a holistic rubric is presented for the first task and an analytic rubric for the second. It should

be noted that either a holistic or an analytic rubric could have been designed for either task.

Example 1:
Subject - Mathematics
Grade Level(s) - Upper Elementary

Mr. Harris, a fourth-grade teacher, is planning a unit on the topic of data analysis, focusing primarily on the skills of estimation and interpretation of graphs. Specifically, at the end of this unit, he wants to be able to assess his students' mastery of the following instructional objectives:

- Students will properly interpret a bar graph.
- Students will accurately estimate values from within a bar graph. (step 1)

Since the purpose of his performance task is summative in nature - the results will be incorporated into the students' grades, he decides to develop a holistic rubric. He identifies the following four attributes on which to focus his rubric: estimation, mathematical computation, conclusions, and communication of explanations (steps 2 & 3). Finally, he begins drafting descriptions of the various levels of performance for the observable attributes (steps 4 & 5). The final rubric for his task appears in Table 4.

Table 4:
Math Performance Task – Scoring Rubric
Data Analysis

Name _____		Date _____
<u>Score</u>	<u>Description</u>	
4	Makes accurate estimations. Uses appropriate mathematical operations with no mistakes. Draws logical conclusions supported by graph. Sound explanations of thinking.	
3	Makes good estimations. Uses appropriate mathematical operations with few mistakes. Draws logical conclusions supported by graph. Good explanations of thinking.	
2	Attempts estimations, although many inaccurate. Uses inappropriate mathematical operations, but with no mistakes. Draws conclusions not supported by graph. Offers little explanation.	
1	Makes inaccurate estimations. Uses inappropriate mathematical operations. Draws no conclusions related to graph. Offers no explanations of thinking.	
0	No response/task not attempted.	

Example 2:
Subjects - Social Studies; Probability & Statistics
Grade Level(s) - 9 - 12

Mrs. Wolfe is a high school American government teacher. She is beginning a unit on the electoral process and knows from past years that her students sometimes have difficulty with the concepts of sampling and election polling. She decides to give her students a performance assessment so they can

demonstrate their levels of understanding of these concepts. The main idea that she wants to focus on is that samples (surveys) can accurately predict the viewpoints of an entire population. Specifically, she wants to be able to assess her students on the following instructional objectives:

- Students will collect data using appropriate methods.
- Students will accurately analyze and summarize their data.
- Students will effectively communicate their results. (step 1)

Since the purpose of this performance task is formative in nature, she decides to develop an analytic rubric focusing on the following attributes: sampling technique, data collection, statistical analyses, and communication of results (steps 2 & 3). She drafts descriptions of the various levels of performance for the observable attributes (steps 4 & 5). The final rubric for this task appears in Table 5.

Table 5: <i>Performance Task – Scoring Rubric</i> <i>Population Sampling</i>					
Name _____			Date _____		
	Beginning 1	Developing 2	Accomplished 3	Exemplary 4	Score
Sampling Technique	Inappropriate sampling technique used	Appropriate technique used to select sample; major errors in execution	Appropriate technique used to select sample; minor errors in execution	Appropriate technique used to select sample; no errors in procedures	
Survey/ Interview Questions	Inappropriate questions asked to gather needed information	Few pertinent questions asked; data on sample is inadequate	Most pertinent questions asked; data on sample is adequate	All pertinent questions asked; data on sample is complete	
Statistical Analyses	No attempt at summarizing collected data	Attempts analysis of data, but inappropriate procedures	Proper analytical procedures used, but analysis incomplete	All proper analytical procedures used to summarize data	
Communication of Results	Communication of results is incomplete, unorganized, and difficult to follow	Communicates some important information; not organized well enough to support decision	Communicates most of important information; shows support for decision	Communication of results is very thorough; shows insight into how data predicted outcome	
Total Score = _____					

Resources for Rubrics on the Web

The following is just a partial list of some Web resources for information about and samples of scoring rubrics.

- "Scoring Rubrics: What, When, & How?" (<http://ericae.net/pare/getvn.asp?v=7&n=3>). This article appears in Practical Assessment. Research. & Evaluation and is authored by Barbara M. Moskal.

The article discusses what rubrics are, and distinguishes between holistic and analytic types. Examples and additional resources are provided.

- "Performance Assessment-Scoring" (<http://www.pgcps.pg.k12.md.us/~elc/scoringtasks.html>). Staff in the Prince George's County (MD) Public Schools have developed a series of pages that provide descriptions of the steps involved in the design of performance tasks. This particular page provides several rubric samples.
- "Rubrics from the Staff Room for Ontario Teachers" (<http://www.odyssey.on.ca/~elaine.coxon/rubrics.htm>) This site is a collection of literally hundreds of teacher-developed rubrics for scoring performance tasks. The rubrics are categorized by subject area and type of task. This is a fantastic resource...check it out!
- "Rubistar Rubric Generator" (<http://rubistar.4teachers.org/>)
- "Teacher Rubric Maker" (http://www.teach-nology.com/web_tools/rubrics/) These two sites house Web-based rubric generators for teachers. Teachers can customize their own rubrics based on templates on each site. In both cases, rubric templates are organized by subject area and/or type of performance task. These are wonderful resources for teachers!

References

- Airasian, P. W. (2000). *Assessment in the classroom: A concise approach* (2nd ed.). Boston: McGraw-Hill.
- Airasian, P. W. (2001). *Classroom assessment: Concepts and applications* (4th ed.). Boston: McGraw-Hill.
- Chase, C. I. (1999). *Contemporary assessment for educators*. New York: Longman.
- Mertler, C. A. (2001). Using performance assessment in your classroom. Unpublished manuscript, Bowling Green State University.
- Montgomery, K. (2001). *Authentic assessment: A guide for elementary teachers*. New York: Longman.
- Moskal, B. M. (2000). Scoring rubrics: what, when, and how?. *Practical Assessment, Research, & Evaluation*, 7(3). Available online: <http://ericae.net/pare/getvn.asp?v=7&n=3>
- Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.
- Tombari, M. & Borich, G. (1999). *Authentic assessment in the classroom: Applications and practice*. Upper Saddle River, NJ: Merrill.
- Trice, A. D. (2000). *A handbook of classroom assessment*. New York: Longman.

Contact information

Craig A. Mertler
Educational Foundations & Inquiry Program
College of Education & Human Development
Bowling Green State University
Bowling Green, OH 43403

mertler@bgnet.bgsu.edu

Phone: 419-372-9357 Fax: 419-372-8265

Descriptors: *Rubrics; Scoring; *Student Evaluation; *Test Construction; *Evaluation Methods; Grades; Grading; *Scoring